
AFRL-AFOSR-JP-TR-2018-0022

# Novel numerical methods for optimal control problems involving fractional-order differential equations

song wang
CURTIN UNIVERSITY OF TECHNOLOGY

03/14/2018
Final Report

DISTRIBUTION A: Distribution approved for public release.

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ IOA
Arlington, Virginia 22203
Air Force Materiel Command

# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 0704-0188*

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
**PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

| 1. REPORT DATE *(DD-MM-YYYY)* | 2. REPORT TYPE | 3. DATES COVERED *(From - To)* |
|---|---|---|
| 12/15/2017 | Final report | Sept 2015 - Sept 2017 |

**4. TITLE AND SUBTITLE**
Novel numerical methods for optimal control problems involving fractional-order differential equations

**5a. CONTRACT NUMBER**
FA2386-15-1-4095

**5b. GRANT NUMBER**
BAA-AFOSR-2014-0001

**5c. PROGRAM ELEMENT NUMBER**

**6. AUTHOR(S)**
Wang, Song

**5d. PROJECT NUMBER**

**5e. TASK NUMBER**

**5f. WORK UNIT NUMBER**

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**
Curtin University of Technology
Kent Street, Bentley WA6102
Australia

**8. PERFORMING ORGANIZATION REPORT NUMBER**

**9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)**
Asian Office of Aerospace of R&D, 7-23-17 Roppongi, Minatu-KU, Tokyo, Japan
SUSAN FULLER
PROCUREMENT ANALYST
E-mail: susan.fuller@us.af.mil
Phone: 703-696-8523

**10. SPONSOR/MONITOR'S ACRONYM(S)**

**11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

**12. DISTRIBUTION/AVAILABILITY STATEMENT**
Papers attached in this report have either been published in or submitted for publication to an international journal.

**13. SUPPLEMENTARY NOTES**

**14. ABSTRACT**
This final report summarizes the activities carried out in the period of this project in which we have developed various novel and efficient numerical methods for solving optimal control and optimization problems with fractional-order differential equation constraints. These methods include discretization schemes for fractional differential equations and algorithms for solving the discretized optimal control/optimization problems. Papers and reports containing the theoretical and numerical results are attached in this final report.

**15. SUBJECT TERMS**
Optimal control, stochastic optimal control, dynamic optimization, fractional differential equations, Caputo's fractional derivative, discretization schemes, gradient based optimization algorithms, control and financial engineering.

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | Song Wang |
| | | | | | 19b. TELEPHONE NUMBER *(Include area code)* |
| U | U | U | UU | 115 | +61-8-9266 2396 |

# Novel numerical methods for optimal control and optimization problems involving fractional-order differential equations

## US Airforce Project 15IOA095

**Principal Investigators: Song Wang and Volker Rehbock**

**Department of Mathematics & Statistics**

**Curtin University**

### Aims of the project

The aim of this project is to develop numerical solution methods for optimal control problems which are subject to systems of fractional differential equations. These systems yield more accurate representations of many real world systems and can incorporate a more global view of the system state. Amongst other advantages, this allows modellers to include features such as memory effects in either space or time. As fractional order systems require quite distinct numerical solution methods, it is a major task to develop numerical methods for both optimal open and closed loop control problems. Our aim is to develop effective and efficient numerical methods for the construction of solutions to fractional order optimal control problems. As very few, if any, such methods currently exist, our work will enable future researchers and practitioners to address and solve practically important optimal control and optimization problems involving fractional-order differential equations.

### Activities and achievements within this project

This project officially started on 23 September 2015 and ended on 22 September 2017. However, our investigation and research activities in the area started early 2015 since we first proposed the project. The research associate, Dr Wen den Hollander started her employment on 24 October 2015. The past two year are very fruitful years during which we have studies various optimal control and optimization problems and their applications arising in both control engineering and financial engineering. Various research activities have been supported by the project in the as outlined below.

1. A $2^{nd}$-order one-point numerical integration scheme has been developed and analysed for solving fractional dynamical systems which is an integral part of optimal control problem. Efficient and accurate numerical methods, essential for solving fractional optional control problems, are scarce in the open literature. In this paper we propose an efficient and easy-to-implement numerical method for an α-th order ordinary differential equation when $\alpha \in (0,1)$, based on a one-point quadrature rule. The quadrature point in each sub-interval of a given partition with mesh size $h$ is chosen judiciously so that the degree of accuracy of the quadrature rule is 2 in the presence of the singular integral kernel. The resulting time-stepping method can be regarded as the counterpart for fractional ODEs of the well-known mid-point method for the 1st-

**DISTRIBUTION A. Approved for public release: distribution unlimited.**

order ODEs. We show that the global error in a numerical solution generated by this method is of the 2nd-order accuracy, independently of α. An extension of this method to dynamical systems involved in optimal control problems has been discussed. Numerical results are presented to demonstrate that the computed rates of convergence match the theoretical one very well and that our method is much more accurate than a well-known one-step method when α is small.

A research paper containing the developed theoretical and numerical results has been published in an international journal.

2. A numerical algorithm combining a generalization of the algorithm in Item 2 above to a system of equations and a gradient-based method is developed for solving general fractional optimal open-loop control problems ($\alpha \in (0,1)$) with multiple states and control variables. This algorithm has the merit that it has a 2nd-order convergence rate and is computationally efficient, and thus can handle large-scale fractional optimal control problems. A gradient formula has been developed which forms the basis of the numerical method for multiple state and control problems. Convergence of the method has been proven. The combined method has been coded using Matlab programming language and extensive numerical experiments have been conducted to demonstrate the performance of the method using optimal control problems with multiple states and controls. The numerical results show that the numerical scheme developed in this project is able to solve fractional optimal control problems of practical significance.

   A research paper containing the detailed description of the method and numerical experimental results has been submitted to an international journal for publication.

3. A 2nd-order finite-difference method for a fractional differential complementarity (variational inequality) problem of order $\alpha \in (1,2)$ arising from the stochastic optimal feedback (closed loop) control in financial engineering. In this work we have designed the finite-difference method for solving the 2nd-order fractional partial differential equation and showed that the truncation error of the method is of 2nd-order. Numerical experiments have been performed to demonstrate the accuracy and efficiency of the method. Dr. Song Wang presented the results as a plenary speaker at the 6th Conference on Numerical Analysis & Applications held in June, 2016 in Lozenets, Bulgaria.

   Two research papers have been in published respectively in an edited volume of Lecture Notes in Computer Science and an international journal.

4. Numerical solution of a high-dimensional Hamilton-Jacobi-Bellman (HJB) equation arising from an optimal control feedback problem in engineering. In this paper we propose a combination of a penalty method and a finite volume scheme for a four-dimensional time-dependent (HJB) equation arising from a stochastic optimal control

problem in pricing financial options with proportional transaction costs and stochastic volatility. The HJB equation is first approximated by a nonlinear differential equation containing penalty terms. A finite volume method along with an upwind technique is then developed for the spatial discrtization of the nonlinear penalty equation. We show that the coefficient matrix of the discretized system is an M-matrix. An iterative method is proposed for solving the nonlinear algebraic system and a convergence theory is established for the iterative method. Numerical experiments are performed using a non-trivial model pricing problem and the numerical results demonstrate the usefulness of the proposed method.

5. During the period of this project, S. Wang has also in-kind contributions towards the development of efficient numerical methods for the optimal control of robots.

**Use of funds**

The funds have mostly been used for the employment of the research associate, Dr. W. den Hollande. Dr. S. Wang's travels to the Bulgarian conference to deliver his plenary address was also partially supported by the project. Dr. S. Wang has also travelled to HK in December 2017, supported by Curtin University and this project, to deliver an invited talk entitled 'Numerical solution of fractional optimal control problems' at 'The Workshop on Variational Analysis & Stochastic Optimization' organized by HK Polytechnic University.

**Publications and reports within this project**

**(W. Li is the maiden name of W. den Hollander.)**


1. W. Li, S. Wang, V. Rehbock, Numerical solution of fractional optimal control, submitted for publication.
2. W. Li, S. Wang, V. Rehbock, A 2$^{nd}$-order one-point numerical integration scheme for fractional ordinary differential equations, Numerical Algebra, Control & Optimization, Vol.7, No.3, 273-287 (2017).
3. W. Chen, S. Wang, A 2nd-Order FDM for a 2D Fractional Black-Scholes Equation. In: Dimov I., Faragó I., Vulkov L. (eds) Numerical Analysis and Its Applications. NAA 2016. Lecture Notes in Computer Science, Vol. 10187. Springer, Cham, 46-57, (2017).
4. W. Chen, S. Wang, A power penalty method for a 2D fractional partial differential linear complementarity problem governing two-asset American options pricing, Appl. Math. Comp. Vol.305, 174-187 (2017)
5. W. Li, S. Wang, Pricing European options with proportional transaction costs and stochastic volatility using a penalty approach and a finite volume scheme, Computer & Mathematics with Applications, Vol.73, 2454-2469 (2017).


**DISTRIBUTION A. Approved for public release: distribution unlimited.**

6. M. Tan, L.S. Jennings, S. Wang, Analysing human periodic walking at different speeds using parametrization enhancing transform in dynamic optimization, *Pacific Journal of Optimization*, Vol.12, 557-586 (2016).

**Attachments: Papers and reports listed above.**

**DISTRIBUTION A. Approved for public release: distribution unlimited.**

# Numerical solution of fractional optimal control *†

Wen Li, Song Wang and Volker Rehbock

Department of Mathematics & Statistics
Curtin University, GPO Box U1987, Perth WA6845, Australia
wen.li@curtin.edu.au; song.wang@curtin.edu.au
V.Rehbock@curtin.edu.au

**Abstract**

This paper presents a numerical algorithm for solving a class of optimal control problems with a dynamic system containing fractional differential equations. We first propose a robust 2nd-order numerical integration scheme for the fractional system, based a set of judiciously chosen quadrature points. The objective is approximated by the trapezoidal rule. We then apply a gradient-based optimization method to solve the discretized optimal control problem. Formulas for calculating the gradients with respect to the unknown discrete control values are derived. Computational results demonstrate that the proposed method is able to generate good numerical approximations for optimal problems with multiple state and control variables. The results also show that the method is robust with respect to the fractional orders of derivatives involved in the dynamics.

## 1 Introduction

A fractional order optimal control problem (FOCP) involves dynamics which are described by fractional differential equations. In the last decade, fractional order optimal control problems have arisen in many fields such as mathematics, engineering, biology, economics, finance and management. Various methods have been developed for solving these problems (see, for example,[1, 2, 3, 4, 5, 7, 12, 18, 19, 25, 26, 32, 37]). In [1, 2], Agrawal extended the classical control theory to fractional dynamic systems and derived fractional Euler-Lagrange equations for FOCPs. These equations give the necessary conditions of optimality for unconstrained FOCPs. The fractional Euler-Lagrange equations have been solved numerically in [1, 3, 5] where the performance index is assumed to be a quadratic function. Based on the work of [1, 2], Singha and Nahak [36] derived necessary optimality conditions for a class of FOCP, where the dynamical constraints comprise a combination of classical and fractional derivatives. In [4, 7, 12, 18, 19, 26, 32, 37]), the authors solved

*This work is supported by the AOARD Project # 15IOA095 from the US Air Force.
†Submitted to an international journal for publication.

**DISTRIBUTION A. Approved for public release: distribution unlimited.**

the FOCPs directly without reference to the necessary optimality conditions of the continuous problem. Tricand and Chen [37] converted FOCPs into a general, rational form of optimal control problem by a rational approximation method. In [4, 7, 12, 18, 19, 26, 32]) the authors used polynomial approximations of the state and control to solve an FOCP. In these methods, they first derived an operational matrix for the fractional derivatives based on the polynomial approximation. Then, the system of equations derived from the dynamic constraints was adjoined to the performance index. By deriving the necessary conditions for the optimality of the performance index, the given FOCP reduces to a problem of solving a system of algebraic equations which can be solved by an iterative method. Bernstein polynomials [4, 32], Jacobi polynomials [12, 18], Legendre polynomials [26, 19] and Chebyshev polynomials [6] have been used in these papers.

Although many researchers have studied FOCPs, most of them considered only one-dimensional FOCPs involving one state variable and one control variable. Recently, Alipour et al. [4] and Bhrawy et al. [7] developed numerical schemes for multi-dimensional FOCPs. In [4], the authors considered a FOCP in which the performance index and the constraint conditions of fractional differential equations are polynomial functions of the state and control variables. Bhrawy et al. [7] solved a multi-dimensional FOCP with a quadratic performance index and linear fractional dynamic constraints. Both of these papers used polynomial approximation methods for solving the FOCPs. In [4], Alipour et al. used Bernstein polynomials, whereas Bhrawy et al.[7] used orthonormal Legendre polynomials. It is well known that approximation of the solution to a differential equation by high-order polynomials often results in ill-conditioned algebraic systems and numerical instability. To our best knowledge, there are no numerical methods in the open literature for general FOCPs with multiple states and controls which are comparable to popular existing numerical methods for conventional optimal control problems.

There are two commonly used definitions of a fractional derivative: the Riemann-Liouville and the Caputo fractional derivative representations [34]. In the paper, we use the Caputo fractional derivative which is defined as follows.

**Definition 1.1** *Assume that $y(t)$ is differentiable on $[0,\infty)$ for a positive constant $T$ and $0<\beta<1$. The Caputo derivative of order $\beta$ of the function $y(t)$ is defined as*

$$ {}_0D_t^{\beta} y(t) = \frac{1}{\Gamma(1-\beta)} \int_0^t \frac{y'(\tau)}{(t-\tau)^{\beta}}\, d\tau $$

*for $t>0$, where $\Gamma(\cdot)$ denotes the Gamma function.*

In what follows, we present a new direct numerical method for a general multi-dimensional FOCPs. The aim is to present a tractable method which can be applied to many FOCPs of practical significance. The general problem considered in the paper is described as follows.

$$ \min_{u\in\mathcal{U}} \quad F(u) = \int_0^T L(t,x(t),u(t))\, dt + S(x(T)), \tag{1.1} $$

$$ \text{subject to} \quad \begin{cases} {}_0D_t^{\alpha} x(t) = f(t,x(t),u(t)), \quad t\in(0,T], \\ x(0) = x^0, \end{cases} \tag{1.2} $$

$$ g(u(t)) \le 0,\ t\in(0,T], \tag{1.3} $$

**DISTRIBUTION A. Approved for public release: distribution unlimited.**

where $x(t) = (x_1(t), x_2(t), \ldots, x_n(t))^\top \in \mathbb{R}^n$ and $u(t) = (u_1(t), u_2(t), \ldots, u_m(t))^\top \in \mathbb{R}^m$ are the state and control variables for some positive integers $n$ and $m$ respectively, $T > 0$ is a fixed constant, $f = (f_1, f_2, \ldots, f_n)^\top$, $L$, $S$ and $g = (g_1, g_2, ..., g_p)^\top$ are known functions for a positive integer $p$, $x^0 \in \mathbb{R}^n$ is a given vector, $\mathcal{U} \subset \mathbb{R}^m$ is the set all bounded piecewise continuous functions on $[0, T]$, and

$$ {}_0D_t^\alpha x(t) = ({}_0D_t^{\alpha_1} x_1(t), {}_0D_t^{\alpha_2} x_2(t), \ldots, {}_0D_t^{\alpha_n} x_n(t))^\top $$

with ${}_0D_t^{\alpha_i} x_i(t)$ denoting Caputo's $\alpha_i$-th derivative of $x_i(t)$ defined in Definition 1.1. In the literature, almost all papers on FOCPs only consider a fractional order $\alpha \in [0.5, 1)$. In this paper, we consider FOCPs for all $\alpha_i \in (0, 1)$ , $i = 1, 2, \ldots, n$.

To solve (1.1)–(1.3) numerically, we need to first introduce an approximation scheme for the system of fractional differential equations (1.2). In the open literature, there are a number of different methods for solving the initial value problem (1.2). See, for example, [14, 15, 16, 17, 9, 21, 22, 23, 24, 27, 28, 31]. However, none of these methods have a satisfactory rate of convergence when $\alpha$ is close to zero. In our recent work [30], we proposed a one-step 2nd-order numerical integration scheme for solving a scalar fractional differential equation based on a one-point quadrature rule with a judiciously chosen point in each mesh subinterval. This one-step numerical integration scheme has a 2nd-order rate of convergence which is independent of $\alpha$. It is also easy to implement and computationally inexpensive. In this paper, we will first extend this method to the system (1.2).

The rest of the paper is organized as the follows. In Section 2, we first approximate the constrained problem (1.1)–(1.3) by an unconstrained one using a well-known penalty approach. Then we convert (1.2) to an equivalent system of Volterra integral equations. In Section 3, we propose a discrete approximation of the objective function and then derive an explicit scheme for the Volterra integral equations based on a Taylor expansion. In Section 4, we derive a formula for calculating the gradient of the discretized objective with respect to the decision variables. Finally, we propose a gradient-based algorithm for the problem on the basis of this gradient formula. In Section 5, numerical examples are presented to demonstrate the accuracy and effectiveness of the proposed method. Section 6 concludes the paper.

# 2 Preliminaries

We first make the following assumptions on the given functions in (1.1)–(1.3):

A1. $f$ is twice continuously differentiable with respect to all its arguments.

A2. $L$ is continuously differentiable in $x$ and $u$.

A3. $S$ and $g$ are continuously differentiable with respect to $x$ and $u$ respectively.

Clearly, (1.1)–(1.3) is a constrained optimal control problem. We first approximate the problem by the following unconstrained optimal control problem using a penalty

**DISTRIBUTION A. Approved for public release: distribution unlimited.**

approach.

$$\min_{u\in\mathcal{U}} \quad \hat{F}(u) := F(u) + \lambda \sum_{j=1}^{p} \int_0^T [g_j(u(t))]_+^2 dt \tag{2.1}$$

$$\text{subject to} \quad \begin{cases} {}_0D_t^\alpha x(t) = f(t, x(t), u(t)), & t \in (0, T], \\ x(0) = x^0, \end{cases} \tag{2.2}$$

where $[z]_+ = \max\{0, z\}$ and $\lambda > 1$ is the penalty constant. This penalty approach has been used extensively in optimization and conventional optimal control [10, 11, 20, 29, 33, 35, 38] and it has been shown that this penalty method is exact in [13, 38].

Since the penalty term in the integrand of (2.1) is smooth, it can be combined with the original objective integrand $L$ to form a new integrand which is still continuously differentiable in $x$ and $u$. Therefore, we may rewrite the penalized problem (2.1)-(2.2) as the following general unconstrained form:

$$\min_{u\in\mathcal{U}} \quad F(u) \tag{2.3}$$

$$\text{subject to} \quad \begin{cases} {}_0D_t^\alpha x(t) = f(t, x(t), u(t)), & t \in (0, T], \\ x(0) = x^0, \end{cases} \tag{2.4}$$

where $x(t), u(t), f, L, S, x^0$ and ${}_0D_t^\alpha x(t)$ are as defined before and $F$ now contains the penalized constraints.

Using Definition 1.1, one can show the following initial value problem is equivalent to a Volterra integral equation as given in the following lemma.

**Lemma 2.1** *Let $\beta \in (0, 1)$ be a constant and $\phi(t, y(t))$ a continuous function. Then the initial value problem*

$$\begin{cases} {}_0D_t^\beta y(t) = \phi(t, y(t)), & t \in (0, T], \\ y(0) = y_0 \end{cases}$$

*is equivalent to the following Volterra integral equation:*

$$y(t) = y_0 + \frac{1}{\Gamma(\beta)} \int_0^t (t-\tau)^{\beta-1} \phi(\tau, y(\tau))\, d\tau, \quad \beta \in (0, 1),$$

*for $t > 0$, where $y_0$ is a given initial condition.*

PROOF. The proof can be found in [8] and is therefore omitted. □

Using Lemma 2.1, we rewrite (2.3)-(2.4) in the following optimal control problem:

$$\min_{u\in\mathcal{U}} \quad F(u) \tag{2.5}$$

$$\text{subject to} \quad x_i(t) = x_i^0 + \frac{1}{\Gamma(\alpha_i)} \int_0^t (t-\tau)^{\alpha_i - 1} f_i(\tau, x(\tau), u(\tau))\, d\tau, \tag{2.6}$$
$$t \in (0, T],\ i = 1, 2, ..., n.$$

**DISTRIBUTION A. Approved for public release: distribution unlimited.**

# 3 Discretization of (2.5) and (2.6)

In this section, we propose an algorithm for the numerical solution of (2.5)–(2.6).

Let $N$ be a given positive integer. We divide $[0,T]$ into a uniform mesh with the mesh points $t_j = jh$ for $j = 0, 1, \ldots, N$, where $h = T/N$. Using this partition, we approximate the objective $F$ in (2.5) with the trapezoidal rule as follows.

$$\begin{aligned} F(u) \approx & \frac{1}{2}L(t_0, x(t_0), u(t_0))h + \sum_{j=1}^{N-1} L(t_j, x(t_j), u(t_j))h \\ & + \frac{1}{2}L(t_N, x(t_N), u(t_N))h + S(x(t_N)). \end{aligned} \tag{3.1}$$

By (2.6), for each $i = 1, 2, ..., n$ and $j = 1, 2, \ldots, N$, we have

$$\begin{aligned} x_i(t_j) &= x_i^0 + \frac{1}{\Gamma(\alpha_i)} \int_0^{t_j} (t_j - \tau)^{\alpha_i - 1} f_i(\tau, x(\tau), u(\tau))\, d\tau \\ &= x_i^0 + \frac{1}{\Gamma(\alpha_i)} \int_0^{jh} (jh - \tau)^{\alpha_i - 1} f_i(\tau, x(\tau), u(\tau))\, d\tau \\ &= x_i^0 + \frac{1}{\Gamma(\alpha_i)} \sum_{k=1}^{j} \int_{(k-1)h}^{kh} (jh - \tau)^{\alpha_i - 1} f_i(\tau, x(\tau), u(\tau)) d\tau. \end{aligned} \tag{3.2}$$

We now consider an approximation for the integral on the right hand side of (3.2). By Assumption A1, $f_i(t, x(t), u(t))$ is twice continuously differentiable with respect to $t$, $x$ and $u$. Thus, for $k = 1, 2, \ldots, j$, we use Taylor's theorem for $f_i(\tau, x(\tau), u(\tau))$ at any point $\tau_{jk}^i \in ((k-1)h, kh)$ to yield

$$f_i(\tau, x(\tau), u(\tau)) = f_i(\tau_{jk}^i, x(\tau_{jk}^i), u(\tau_{jk}^i)) + K_{jk}^i(\tau - \tau_{jk}^i) + c_{jk}^i(\tau - \tau_{jk}^i)^2, \tag{3.3}$$

where $c_{jk}^i$ is the coefficient of the reminder of the expansion and

$$\begin{aligned} K_{jk}^i = & \frac{\partial f_i}{\partial \tau}\Big|_{(\tau_{jk}^i, x(\tau_{jk}^i), u(\tau_{jk}^i))} + \sum_{l=1}^{n} \frac{\partial f_i}{\partial x_l}\Big|_{(\tau_{jk}^i, x(\tau_{jk}^i), u(\tau_{jk}^i))} \frac{\partial x_l}{\partial \tau}|_{(\tau_{jk}^i)} \\ & + \sum_{r=1}^{m} \frac{\partial f_i}{\partial u_r}\Big|_{(\tau_{jk}^i, x(\tau_{jk}^i), u(\tau_{jk}^i))} \frac{\partial u_l}{\partial \tau}\Big|_{(\tau_{jk}^i)} \end{aligned}$$

Therefore, replacing $f_i(\tau, x(\tau), u(\tau))$ in the integrand of the last term in (3.2) with the

**DISTRIBUTION A. Approved for public release: distribution unlimited.**

RHS of (3.3), we have, for any $k = 1, ..., j$,

$$\begin{aligned}
&\frac{1}{\Gamma(\alpha_i)} \int_{(k-1)h}^{kh} (jh-\tau)^{\alpha_i-1} f_i(\tau, x(\tau), u(\tau)) d\tau \\
&\quad = \frac{1}{\Gamma(\alpha_i)} \int_{(k-1)h}^{kh} (jh-\tau)^{\alpha_i-1} [f_i(\tau_{jk}^i, x(\tau_{jk}^i), u(\tau_{jk}^i)) + K_{jk}^i(\tau - \tau_{jk}^i)] d\tau + R_{jk}^i \\
&\quad = \frac{1}{\Gamma(\alpha_i)} f_i\left(\tau_{jk}^i, x(\tau_{jk}^i), u(\tau_{jk}^i)\right) \left[\frac{(jh-(k-1)h)^{\alpha_i}}{\alpha_i} - \frac{(jh-kh)^{\alpha_i}}{\alpha_i}\right] \\
&\qquad + \frac{K_{jk}^i}{\Gamma(\alpha_i)} \int_{(k-1)h}^{kh} (jh-\tau)^{\alpha_i-1} (\tau - \tau_{jk}^i)\, d\tau + R_{jk}^i \\
&\quad = \frac{h^{\alpha_i}}{\Gamma(\alpha_i+1)} f_i(\tau_{jk}^i, x(\tau_{jk}^i), u(\tau_{jk}^i))[(j-k+1)^{\alpha_i} - (j-k)^{\alpha_i}] \\
&\qquad + \frac{K_{jk}^i}{\Gamma(\alpha_i)} \int_{(k-1)h}^{kh} (jh-\tau)^{\alpha_i-1} (\tau - \tau_{jk}^i) d\tau + R_{jk}^i, \qquad (3.4)
\end{aligned}$$

where $R_{jk}^i = \dfrac{1}{\Gamma(\alpha_i)} \displaystyle\int_{(k-1)h}^{kh} (jh-\tau)^{\alpha_i-1} c_{jk}^i (\tau - \tau_{jk}^i)^2 d\tau$.

We now consider the choice of $\tau_{jk}^i$. From (3.4) it is clear that $\tau_{jk}^i$ should be chosen such that the second term becomes zero so that the truncation error in (3.4) is $R_{jk}^i$. This choice of $\tau_{jk}^i$ is given in the following theorem.

**Theorem 3.1** *For any given $j \in \{1, 2, \ldots, N\}$ and $k \in \{1, 2 \ldots, j\}$, the unique solution to*

$$\int_{(k-1)h}^{kh} (jh-\tau)^{\alpha_i-1} (\tau - \tau_{jk}^i) d\tau = 0$$

*is given by*

$$\tau_{jk}^i = h \frac{[(j-k+1)^{\alpha_i+1} - (j-k)^{\alpha_i+1}] + (\alpha_i+1)[(j-k+1)^{\alpha_i}(k-1) - (j-k)^{\alpha_i} k]}{(\alpha_i+1)[(j-k+1)^{\alpha_i} - (j-k)^{\alpha_i}]}. \qquad (3.5)$$

*Furthermore, $(k-1)h < \tau_{jk}^i < kh$.*

PROOF. See the proof of Theorem 2.1 in [30]. □

Substituting the expression for $\tau_{jk}^i$ in (3.5) into (3.4) and combining the resulting expression with (3.2), we have the following representation for $x_i(t_j)$.

$$x_i(t_j) = x_i^0 + \frac{h^{\alpha_i}}{\Gamma(\alpha_i+1)} \sum_{k=1}^{j} f_i(\tau_{jk}^i, x(\tau_{jk}^i), u(\tau_{jk}^i))[(j-k+1)^{\alpha_i} - (j-k)^{\alpha_i}] + R_j^i, \quad (3.6)$$

for $j = 1, 2, \ldots, N$, where $\tau_{jk}^i$ is given in (3.5) for $k = 1, 2, \ldots, j$ and $R_j^i = \sum_{k=1}^{j} R_{jk}^i$. Omitting the remainder $R_j^i$ in (3.6), we have an equation approximating (3.2) which has the truncation error $R_j^i$. An upper bound for $R_j^i$ is given in the following theorem.

**DISTRIBUTION A. Approved for public release: distribution unlimited.**

**Theorem 3.2** *Let Assumption A1 be fulfilled. Then the following estimate holds:*

$$|R_j^i| \le Ch^2,$$

*where $C$ denotes a positive constant independent of $h$.*

PROOF. See the proof of Theorem 2.2 in [30]. □

From (3.6) it is clear that to compute $x_i(t_j)$, we need to calculate $f_i(\tau_{jk}^i, x(\tau_{jk}^i), u(\tau_{jk}^i))$, where $x(\tau_{jk}^i) = (x_1(\tau_{jk}^i), x_2(\tau_{jk}^i), \ldots, x_n(\tau_{jk}^i))$, $u(\tau_{jk}^i) = (u_1(\tau_{jk}^i), u_2(\tau_{jk}^i), \ldots, u_m(\tau_{jk}^i))$ and $k = 1, 2, \ldots, j$. However, $x(\tau_{jk}^i), u(\tau_{jk}^i)$ are not available directly from the scheme, although the points $\tau_{jk}^i$ for feasible $i, j$ and $k$ are known. Thus, approximations for $x(\tau_{jk}^i)$ and $u(\tau_{jk}^i)$ need to be determined. Next, we propose a numerical scheme based on a linear interpolation and a Taylor expansion for approximating $x(\tau_{jk}^i)$ and $u(\tau_{jk}^i)$.

For any indices $j$ and $k$ satisfying $1 \le k \le j \le N$, since $\tau_{jk}^i \in (t_{k-1}, t_k)$ by Theorem 3.1, we use the following linear interpolation to approximate $x_i(\tau_{jk}^i)$ and $u_r(\tau_{jk}^i)$:

$$x(\tau_{jk}^i) \approx x(t_{k-1}) + \rho_{jk}^i(x(t_k) - x(t_{k-1})), \tag{3.7}$$

$$u(\tau_{jk}^i) \approx u(t_{k-1}) + \rho_{jk}^i(u(t_k) - u(t_{k-1})), \tag{3.8}$$

where

$$\rho_{jk}^i := \frac{\tau_{jk}^i - t_{k-1}}{h} \in (0, 1). \tag{3.9}$$

The truncation error in the above linear interpolation is of order $\mathcal{O}(h^2)$. Using (3.7) and (3.8), we approximate $f_i(\tau_{jk}^i, x(\tau_{jk}^i), u(\tau_{jk}^i))$ as follows.

$$\begin{aligned} &f_i(\tau_{jk}^i, x(\tau_{jk}^i), u(\tau_{jk}^i)) \\ &\quad \approx f_i\left(\tau_{jk}^i, x(t_{k-1}) + \rho_{jk}^i(x(t_k) - x(t_{k-1})), u(t_{k-1}) + \rho_{jk}^i(u(k_j) - u(t_{k-1}))\right) \end{aligned} \tag{3.10}$$

The truncation error for the above approximation is also of order $\mathcal{O}(h^2)$.

Replacing $f_i(\tau_{jk}^i, x(\tau_{jk}^i), u(\tau_{jk}^i))$ in (3.6) with the RHS of (3.10), we have, up to some terms of order $\mathcal{O}(h^2)$, the following scheme for (2.6):

$$\begin{aligned} x_i(t_j) = x_i^0 + h_{\alpha_i} \sum_{k=1}^{j} \Big[ &f_i\left(\tau_{jk}^i, x(t_{k-1}) + \rho_{jk}^i(x(t_k) - x(t_{k-1})), u(t_{k-1}) + \rho_{jk}^i(u(t_k) - u(t_{k-1}))\right) \\ &\cdot ((j-k+1)^{\alpha_i} - (j-k)^{\alpha_i}) \Big] \end{aligned} \tag{3.11}$$

for $j = 1, 2, \ldots, N$, where $h_{\alpha_i} = \frac{h^{\alpha_i}}{\Gamma(\alpha_i+1)}$ and $\tau_{jk}^i$ is defined by (3.5). Clearly, (3.11) defines a time-stepping scheme for (2.6) with a truncation error of order $\mathcal{O}(h^2)$ because of Theorem 3.2 and the truncation error in (3.10).

The above scheme is implicit as it constitutes a nonlinear system in $x(t_j) = (x_1(t_j), \cdots, x_n(t_j))^\top$. An iterative method such as a Newton's method can be used for solving (3.11). However,

**DISTRIBUTION A. Approved for public release: distribution unlimited.**

it is also possible to define an explicit single step scheme by further approximating the $j$th term in the sum in (3.11) by the following Taylor expansion:

$$\begin{aligned}
&f_i(\tau_{jj}^i, x(t_{j-1}) + \rho_{jj}^i(x(t_j) - x(t_{j-1})), u(t_{j-1}) + \rho_{jj}^i(u(t_j) - u(t_{j-1})))\\
&\quad = f_i(\tau_{jj}^i, x(t_{j-1}), u(t_{j-1}) + \rho_{jj}^i(u(t_j) - u(t_{j-1})))\\
&\quad + \sum_{l=1}^{n} \frac{\partial f_i}{\partial x_l}\Big|_{(\tau_{jj}^i, x(t_{j-1}), u(t_{j-1}) + \rho_{jj}^i(u(t_j) - u(t_{j-1})))} (\rho_{jj}^i(x_l(t_j) - x_l(t_{j-1})))\\
&\quad + \mathcal{O}(h^2).
\end{aligned} \tag{3.12}$$

Thus, combining (3.12) and (3.11) yields

$$\begin{aligned}
x_i(t_j) =& x_i^0 + h_{\alpha_i} \sum_{k=1}^{j-1} \Big[ f_i\left(\tau_{jk}^i, x(t_{k-1}) + \rho_{jk}^i(x(t_k) - x(t_{k-1})), u(t_{k-1}) + \rho_{jk}^i(u(t_k) - u(t_{k-1}))\right)\\
&((j-k+1)^{\alpha_i} - (j-k)^{\alpha_i})\Big] + h_{\alpha_i} f_i\left(\tau_{jj}^i, x(t_{j-1}), u(t_{j-1}) + \rho_{jj}^i(u(t_j) - u(t_{j-1}))\right)\\
&+ h_{\alpha_i} \sum_{l=1}^{n} \Big[ \frac{\partial f_i}{\partial x_l}|_{(\tau_{jj}^i, x(t_{j-1}), u(t_{j-1}) + \rho_{jj}^i(u(t_j) - u(t_{j-1})))} (\rho_{jj}^i(x_l(t_j) - x_l(t_{j-1})))\Big] + \mathcal{O}(h^2).
\end{aligned} \tag{3.13}$$

Let $x^j := (x_1(t_j)), x_2(t_j), \ldots, x_n(t_j))^\top$ and $u^j := (u_1(t_j), u_2(t_j), \cdots, u_m(t_j))^\top$ for $j = 0, 1, ..., N$ with the given initial condition $x^0$. Omitting the truncation error terms of order $\mathcal{O}(h^2)$ and re-organising (3.13), we have the following linear system for $x^j$.

$$B^j(x^{j-1}, u^{j-1}, u^j)x^j = C^j(x^0, x^1, x^2, ..., x^{j-1}, u^0, u^1, ..., u^j), \quad j = 1, 2, \ldots, N. \tag{3.14}$$

where $B^j$ is the $n \times n$ matrix given by

$$B^j = \begin{pmatrix} 1 - b_{11}^j & -b_{12}^j & \ldots & -b_{1n}^j \\ -b_{21}^j & 1 - b_{22}^j & \ldots & -b_{2n}^j \\ \vdots & \vdots & \ddots & \vdots \\ -b_{n1}^j & -b_{n2}^j & \ldots & 1 - b_{nn}^j \end{pmatrix} \tag{3.15}$$

with

$$b_{il}^j = \rho_{jj}^i h_{\alpha_i} \frac{\partial f_i}{\partial x_l}\Big|_{(\tau_{jj}^i, x^{j-1}, u^{j-1} + \rho_{jj}^i(u^j - u^{j-1}))} \tag{3.16}$$

for $i = 1, 2, \ldots, n$, $l = 1, 2, \ldots, n$ and $C^j = (c_1^j, c_2^j, \ldots, c_n^j)^\top$ with

$$\begin{aligned}
c_i^j =& x_i^0 + h_{\alpha_i} \sum_{k=1}^{j-1} \Big[ f_i\left(\tau_{jk}^i, x^{k-1} + \rho_{jk}^i(x^k - x^{k-1}), u^{k-1} + \rho_{jk}^i(u^k - u^{k-1})\right)\\
&((j-k+1)^{\alpha_i} - (j-k)^{\alpha_i}))\Big] + h_{\alpha_i} f_i\left(\tau_{jj}^i, x^{j-1}, u^{j-1} + \rho_{jj}^i(u^j - u^{j-1})\right)\\
&- \sum_{l=1}^{n} x_l^{j-1} b_{il}^j.
\end{aligned} \tag{3.17}$$

It is clear that to calculate $x^j$, we need to solve the system of equations (3.14)-(3.17). Next we show that (3.14)-(3.17) is uniquely solvable when $h$ is sufficiently small.

**DISTRIBUTION A. Approved for public release: distribution unlimited.**

**Theorem 3.3** *The system (3.14)-(3.17) has a unique solution when $h$ is sufficiently small.*

PROOF. We will first show that for $j = 1, 2, \ldots, N$, $B^j$ is a strictly diagonally dominant matrix, i.e.,

$$1 - b_{ii}^j > \sum_{l=1, l\neq i}^{n} |b_{il}^j|, \quad i = 1, 2, ..., n$$

Since

$$1 - b_{ii}^j \geq 1 - |b_{ii}^j|,$$

we only need to show that

$$1 - |b_{ii}^j| > \sum_{l=1, l\neq i}^{n} |b_{il}^j|,$$

or equivalently,

$$\sum_{l=1}^{n} |b_{il}^j| < 1, \quad \forall i.$$

Note that $\rho_{jj}^i \in (0, 1)$ by (3.9) and $h_{\alpha_i} > 0$. We have, from (3.16),

$$|b_{il}^j| < h_{\alpha_i} \left| \frac{\partial f_i}{\partial x_l} |_{(\tau_{jj}^i, x(t_{j-1}), u(t_{j-1}) + \rho_{jj}^i (u(t_j) - u(t_{j-1})))} \right|.$$

Since $f_i$ is twice differentiable in $x$ on $[0, 1]$, $\frac{\partial f_i}{\partial x_l}$ is bounded on $[0, 1]$. Let

$$M = \max_{\substack{1\leq i\leq n \\ 1\leq l\leq n}} \left| \frac{\partial f_i}{\partial x_l} \right|.$$

We have

$$|b_{il}^j| < h_{\alpha_i} M = \frac{h^{\alpha_i}}{\Gamma(\alpha_i + 1)} M.$$

Choose $\bar{h}_i = (\frac{\Gamma(\alpha_i+1)}{nM})^{\frac{1}{\alpha_i}}$ and $\bar{h} = \min_{1\leq i\leq n}\{\bar{h}_i\}$. When $h < \bar{h}$, we have

$$\sum_{l=1}^{n} |b_{il}^j| < \frac{1}{n} n = 1, \quad \forall i.$$

Thus, $B^j$ is a strictly diagonally dominant matrix for all $j$. By the well-known Levy-Desplanques theorem, we conclude that $B^j$ is a non-singular matrix and therefore the system (3.14)-(3.17) has a unique solution. □

Note that, for a given initial condition $x^0$, (3.14) provides a one-step explicit scheme for approximating the solution to (2.6). Introduce $X^j = (x_1^j, x_2^j, ..., x_n^j)^\top$ and $U^j = (u_1^j, u_2^j, ..., u_m^j)^\top$ for $j = 0, 1, ..., N$, and let $X = (X^0, X^1, ..., X^N) \in \mathbb{R}^{n\times(N+1)}$ and $U = (U^0, U^1, ..., U^N) \in \mathbb{R}^{n\times(N+1)}$. Using the approximation schemes defined in (3.1) and (3.14),

**DISTRIBUTION A. Approved for public release: distribution unlimited.**

we pose the following finite-dimensional optimal control problem approximating (2.5)–(2.6):

$$\min_{U\in\mathbb{R}^{m\times(N+1)}} \quad F_h(X,U) := \frac{1}{2}\left(L(t_0,X^0,U^0)+L(t_N,X^N,U^N)\right)h + \sum_{j=1}^{N-1} L(t_j,X^j,U^j)h + S(X^N), \tag{3.18}$$

$$\text{subject to} \quad \begin{cases} B^j(X^{j-1},U^{j-1},U^j)X^j = C^j(X^0,X^1,...,X^{j-1},U^0,U^1,...,U^j), \\ \quad j=1,2,...,N, \\ X^0 = x^0. \end{cases} \tag{3.19}$$

A solution $(X,U)$ to (3.18)–(3.19) is an approximation to a solution $(x(t),u(t))$ of (2.5)–(2.6) at the mesh nodes $t_j, j=0,1,...,N$.

# 4 Solution strategy

In this section, we first determine the gradient of the objective (3.18) with respect to all $U$. We then develop an algorithm for finding approximate solutions to the problem (3.18)-(3.19).

From the definition of $U$ we see that the 1st-order optimality conditions for (3.18) and (3.19) are

$$\frac{\partial F_h}{\partial u_r^j} = 0, \quad r=1,2,\ldots,m, \quad j=0,1,\ldots,N. \tag{4.1}$$

We now determine the LHS of (4.1). From (3.18) and (3.14)–(3.17), we have

$$\begin{aligned}
\frac{\partial F_h}{\partial u_r^0} =& \frac{1}{2}h\frac{\partial L(t_0,X^0,U^0)}{\partial u_r^0} + h\sum_{p=1}^{N-1}\sum_{i=1}^{n}\frac{\partial L(t_p,X^p,U^p)}{\partial x_i^p}\frac{\partial x_i^p}{\partial u_r^0} \\
&+ \frac{1}{2}h\sum_{i=1}^{n}\frac{\partial L(t_N,X^N,U^N))}{\partial x_i^N}\frac{\partial x_i^N}{\partial u_r^0} + \sum_{i=1}^{n}\frac{\partial S(X_N)}{\partial x_i^N}\frac{\partial x_i^N}{\partial u_r^0}, \\
\frac{\partial F_h}{\partial u_r^j} =& h\frac{\partial L(t_j,X^j,U^j)}{\partial u_r^j} + h\sum_{p=1}^{N-1}\sum_{i=1}^{n}\frac{\partial L(t_p,X^p,U^p)}{\partial x_i^p}\frac{\partial x_i^p}{\partial u_r^j} \\
&+ \frac{1}{2}h\sum_{i=1}^{n}\frac{\partial L(t_N,X^N,U^N)}{\partial x_i^N}\frac{\partial x_i^N}{\partial u_r^j} + \sum_{i=1}^{n}\frac{\partial S(X^N)}{\partial x_i^N}\frac{\partial x_i^N}{\partial u_r^j}
\end{aligned}$$

for $j=1,\ldots,N$ and $r=1,2,\ldots,m$.

From (3.19) we see that $\frac{\partial x_i^p}{\partial u_r^j} = 0$ when $p<j$. Thus, the above expression can be

DISTRIBUTION A. Approved for public release: distribution unlimited.

rewritten as

$$\begin{aligned}\frac{\partial F_h}{\partial u_r^0} =&\frac{1}{2}h\frac{\partial L}{\partial u_r}\Big|_{(t_0,X^0,U^0)} + h\sum_{p=1}^{N-1}\sum_{i=1}^{n}\frac{\partial L}{\partial x_i}\Big|_{(t_p,X^p,U^p)}\frac{\partial x_i^p}{\partial u_r^0}\\ &+\frac{1}{2}h\sum_{i=1}^{n}\frac{\partial L}{\partial x_i}\Big|_{(t_N,X^N,U^N)}\frac{\partial x_i^N}{\partial u_r^0} + \sum_{i=1}^{n}\frac{\partial S}{\partial x_i}\Big|_{(X^N)}\frac{\partial x_i^N}{\partial u_r^0}, \qquad (4.2)\\ \frac{\partial F_h}{\partial u_r^j} =&h\frac{\partial L}{\partial u_r}\Big|_{(t_j,X^j,U^j)} + h\sum_{p=j}^{N-1}\sum_{i=1}^{n}\frac{\partial L}{\partial x_i}\Big|_{(t_p,X^p,U^p)}\frac{\partial x_i^p}{\partial u_r^j}\\ &+\frac{1}{2}h\sum_{i=1}^{n}\frac{\partial L}{\partial x_i}\Big|_{(t_N,X^N,U^N)}\frac{\partial x_i^N}{\partial u_r^j} + \sum_{i=1}^{n}\frac{\partial S}{\partial x_i}\Big|_{(X^N)}\frac{\partial x_i^N}{\partial u_r^j}, \quad j=1,2,\ldots,N. \qquad (4.3)\end{aligned}$$

We now need to determine $\frac{\partial x_i^p}{\partial u_r^j}$ for $j = 0, 1, 2, \ldots, N$. By (3.14), we have

$$B^pX^p = C^p, \quad p = 1, 2, \ldots, N.$$

Taking the derivative w.r.t. $u_r^j$ on both sides of the above equation gives

$$\frac{\partial B^p}{\partial u_r^j}X^p + B^p\frac{\partial X^p}{\partial u_r^j} = \frac{\partial C^p}{\partial u_r^j}.$$

Rearranging the above equation, we have

$$B^p\frac{\partial X^p}{\partial u_r^j} = \frac{\partial C^p}{\partial u_r^j} - \frac{\partial B^p}{\partial u_r^j}X^p, \qquad (4.4)$$

where $B^p$ is defined in (3.15), $\frac{\partial X^p}{\partial u_r^j} = (\frac{\partial x_1^p}{\partial u_r^j}, \frac{\partial x_2^p}{\partial u_r^j}, \ldots, \frac{\partial x_n^p}{\partial u_r^j})^\top$,

$$\frac{\partial C^p}{\partial u_r^j} = \Big(\frac{\partial c_1^p}{\partial u_r^j}, \frac{\partial c_2^p}{\partial u_r^j}, \ldots, \frac{\partial c_n^p}{\partial u_r^j}\Big)^\top, \qquad (4.5)$$

and

$$\frac{\partial B^p}{\partial u_r^j} = \begin{pmatrix} -\frac{\partial b_{11}^p}{\partial u_r^j} & -\frac{\partial b_{12}^p}{\partial u_r^j} & \cdots & -\frac{\partial b_{1n}^p}{\partial u_r^j} \\ -\frac{\partial b_{21}^p}{\partial u_r^j} & -\frac{\partial b_{22}^p}{\partial u_r^j} & \cdots & -\frac{\partial b_{2n}^p}{\partial u_r^j} \\ \cdots & \cdots & \cdots & \cdots \\ -\frac{\partial b_{n1}^p}{\partial u_r^j} & -\frac{\partial b_{n2}^p}{\partial u_r^j} & \cdots & -\frac{\partial b_{nn}^p}{\partial u_r^j} \end{pmatrix}. \qquad (4.6)$$

Using (3.16) and (3.17) we see that the entries of the above matrices are given by

$$\begin{aligned}\frac{\partial b_{il}^p}{\partial u_r^j} =&\rho_{pp}^i h_{\alpha_i}\frac{\partial\left(\frac{\partial f_i}{\partial x_l}|_{(\tau_{pp}^i, X^{p-1}, U^{p-1}+\rho_{pp}^i(U^p-U^{p-1}))}\right)}{\partial u_r^j},\\ \frac{\partial c_i^p}{\partial u_r^j} =&\sum_{k=1}^{p-1}h_{\alpha_i}\Bigg(\frac{\partial(f_i(\tau_{pk}^i, X^{k-1}+\rho_{pk}^i(X^k-X^{k-1}), U^{k-1}+\rho_{pk}^i(U^k-U^{k-1})}{\partial u_r^j}\\ &\cdot((p-k+1)^{\alpha_i}-(p-k)^{\alpha_i}))\Bigg)\\ &+h_{\alpha_i}\frac{\partial f_i\left(\tau_{pp}^i, X^{p-1}, U^{p-1}+\rho_{pp}^i(U^p-U^{p-1})\right)}{\partial u_r^j} - \sum_{l=1}^{n}(\frac{\partial x_l^{p-1}}{\partial u_r^j}b_{il}^p + x_l^{p-1}\frac{\partial b_{il}^p}{\partial {u_r}^j})\end{aligned}$$

**DISTRIBUTION A. Approved for public release: distribution unlimited.**

for $i = 1, 2, \ldots, n$, $l = 1, 2, \ldots, n$.

To calculate $\frac{\partial b^p_{il}}{\partial u^j_r}$ and $\frac{\partial c^p_i}{\partial u^j_r}$ in (4.5)–(4.6) for $p = 1, 2, \ldots, N$ and $j = 0, 1, ..., N$, we use the following algorithm:

**Algorithm A**

1. When $j = 0$, if $p = 1$, then

$$\begin{aligned}\frac{\partial b^1_{il}}{\partial u^0_r} &= \rho^i_{11} h_{\alpha_i} \frac{\partial^2 f_i}{\partial x_l \partial u_r}\Big|_{(\tau^i_{11}, X^0, U^0 + \rho^i_{11}(U^1 - U^0))} (1 - \rho^i_{11}),\\ \frac{\partial c^1_i}{\partial u^0_r} &= h_{\alpha_i} \frac{\partial f_i}{\partial u_r}|_{(\tau^i_{11}, X^0, U^0 + \rho^i_{11}(U^1 - U^0)} (1 - \rho^i_{11}) - \sum_{l=1}^{n} (x^0_l \frac{\partial b^1_{il}}{\partial u_r{}^0}).\end{aligned}$$

   If $p > 1$, then

$$\begin{aligned}\frac{\partial b^p_{il}}{\partial u^0_r} =& \rho^i_{pp} h_{\alpha_i} \sum_{q=1}^{n} \Big( \frac{\partial^2 f_i}{\partial x_l \partial x_q}|_{(\tau^i_{pp}, X^{p-1}, U^{p-1} + \rho^i_{pp}(U^p - U^{p-1}))} \frac{\partial x^{p-1}_q}{\partial u^0_r} \Big),\\ \frac{\partial c^p_i}{\partial u^0_r} =& h_{\alpha_i} \frac{\partial f_i}{\partial u_r}|_{(\tau^i_{p1}, X^0 + \rho^i_{p1}(X^1 - X^0), U^0 + \rho^i_{p1}(U^1 - U^0)} (1 - \rho^i_{p1})(p^{\alpha_i} - (p-1)^{\alpha_i})\\ &+ h_{\alpha_i} \Bigg\{ \sum_{k=2}^{p-1} \Bigg[ \sum_{q=1}^{n} \Big( \frac{\partial f_i}{\partial x_q}|_{(\tau^i_{pk}, X^{k-1} + \rho^i_{pk}(X^k - X^{k-1}), U^{k-1} + \rho^i_{pk}(U^k - U^{k-1}))} (1 - \rho^i_{pk}) \frac{\partial x^{k-1}_q}{\partial u^0_r}\\ &+ \frac{\partial f_i}{\partial x_q}|_{(\tau^i_{pk}, X^{k-1} + \rho^i_{pk}(X^k - X^{k-1}), U^{k-1} + \rho^i_{pk}(U^k - U^{k-1}))} \rho^i_{pk} \frac{\partial x^k_q}{\partial u^0_r} \Big)\\ &\Big( (p-k+1)^{\alpha_i} - (p-k)^{\alpha_i} \Big) \Bigg] \Bigg\} + h_{\alpha_i} \Big( \sum_{q=1}^{n} \frac{\partial f_i}{\partial x_q}|_{(\tau^i_{pp}, X^{p-1}, U^{p-1} + \rho^i_{pp}(U^p - U^{p-1}))} \frac{\partial x^{p-1}_q}{\partial u^0_r} \Big)\\ &- \sum_{l=1}^{n} \Big( \frac{\partial x^{p-1}_l}{\partial u^0_r} b^p_{il} + x^{p-1}_l \frac{\partial b^p_{il}}{\partial u^0_r} \Big).\end{aligned}$$

2. For $j = 1, 2, \ldots, N$, if $p < j$, then $\frac{\partial b^p_{il}}{\partial u^j_r} = 0$, $\frac{\partial c^p_i}{\partial u^j_r} = 0$.
   If $p = j$, then

$$\begin{aligned}\frac{\partial b^p_{il}}{\partial u^j_r} &= \rho^i_{jj} h_{\alpha_i} \frac{\partial^2 f_i}{\partial x_l \partial u_r}|_{(\tau^i_{pp}, X^{j-1}, U^{j-1} + \rho^i_{jj}(U^j - U^{j-1}))} \rho^i_{jj},\\ \frac{\partial c^p_i}{\partial u^l_r} &= h_{\alpha_i} \frac{\partial f_i}{\partial u_r}|_{(\tau^i_{jj}, X^{j-1}, u^{j-1} + \rho^i_{pp}(u^j - u^{j-1}))} \rho^i_{jj} - \sum_{l=1}^{n} \Big( x^{j-1}_l \frac{\partial b^j_{il}}{\partial u^j_r} \Big).\end{aligned}$$

**DISTRIBUTION A. Approved for public release: distribution unlimited.**

If $p = j + 1$, then

$$\begin{aligned}
\frac{\partial b_{il}^p}{\partial u_r^j} &= \rho_{pp}^i h_{\alpha_i} \frac{\partial^2 f_i}{\partial x_l \partial u_r}|_{(\tau_{pp}^i, X^{p-1}, U^{p-1}+\rho_{pp}^i(U^p - U^{p-1}))}(1-\rho_{pp}^i) \\
&+ \rho_{pp}^i h_{\alpha_i} \Big( \sum_{q=1}^{n} \frac{\partial^2 f_i}{\partial x_l \partial x_q}|_{(\tau_{pp}^i, X^{p-1}, U^{p-1}+\rho_{pp}^i(U^p - U^{p-1}))} \frac{\partial x_q^{p-1}}{\partial u_r^j} \Big) \\
\frac{\partial c_i^p}{\partial u_r^j} &= h_{\alpha_i} \Bigg[ \frac{\partial f_i}{\partial u_r}|_{(\tau_{pj}^i, X^{j-1}+\rho_{pj}^i(X^j - X^{j-1}), U^{j-1}+\rho_{pj}^i(U^j - U^{j-1})\rho_{pj}^i} \\
&+ \sum_{q=1}^{n} \Big( \frac{\partial f_i}{\partial x_q}|_{(\tau_{pj}^i, X^{j-1}+\rho_{pj}^i(X^j - X^{j-1}), U^{j-1}+\rho_{pj}^i(U^j - U^{j-1}))} \rho_{pj}^i \frac{\partial x_q^{j-1}}{\partial u_r^j} \Big) \\
&\quad (2^{\alpha_i} - 1) \Bigg] + h_{\alpha_i} \Bigg[ \frac{\partial f_i}{\partial u_r}|_{(\tau_{pp}^i, X^{p-1}, U^{p-1}+\rho_{pp}^i(U^p - U^{p-1}))}(1-\rho_{pp}^i) \\
&+ \sum_{q=1}^{n} \Big( \frac{\partial f_i}{\partial x_q}|_{(\tau_{pp}^i, X^{p-1}, U^{p-1}+\rho_{pp}^i(U^p - U^{p-1}))} \frac{\partial x_q^{p-1}}{\partial u_r^j} \Big) \Bigg] \\
&- \sum_{l=1}^{n} \Big( \frac{\partial x_l^{p-1}}{\partial u_r^j} b_{il}^p + x_l^{p-1} \frac{\partial b_{il}^p}{\partial {u_r}^j} \Big)
\end{aligned}$$

If $p > j + 1$, then

$$\begin{aligned}
\frac{\partial b_{il}^p}{\partial u_r^j} &= \rho_{pp}^i h_{\alpha_i} \sum_{q=1}^{n} \frac{\partial^2 f_i}{\partial x_l \partial x_q}|_{(\tau_{pp}^i, X^{p-1}, U^{p-1}+\rho_{pp}^i(U^p - U^{p-1}))} \frac{\partial x_q^{p-1}}{\partial u_r^j} \\
\frac{\partial c_i^p}{\partial u_r^j} &= h_{\alpha_i} \Bigg\{ \sum_{k=j}^{p-1} \Bigg[ \frac{\partial f_i}{\partial u_r}|_{(\tau_{pk}^i, X^{k-1}+\rho_{pk}^i(X^k - X^{k-1}), U^{k-1}+\rho_{pk}^i(U^k - U^{k-1}))} \psi(\rho_{pk}^i) \\
&+ \sum_{q=1}^{n} \Bigg( \frac{\partial f_i}{\partial x_q}|_{(\tau_{pk}^i, X^{k-1}+\rho_{pk}^i(X^k - X^{k-1}), U^{k-1}+\rho_{pk}^i(U^k - U^{k-1}))}(1-\rho_{pk}^i) \frac{\partial x_q^{k-1}}{\partial u_r^j} \\
&\quad + \frac{\partial f_i}{\partial x_q}|_{(\tau_{pk}^i, X^{k-1}+\rho_{pk}^i(X^k - X^{k-1}), U^{k-1}+\rho_{pk}^i(U^k - U^{k-1}))} \rho_{pk}^i \frac{\partial x_q^k}{\partial u_r^j} \Bigg) \\
&\quad (p-k+1)^{\alpha_i} - (p-k)^{\alpha_i)} \Bigg] \Bigg\} \\
&+ h_{\alpha_i} \Bigg( \sum_{q=1}^{n} \Big( \frac{\partial f_i}{\partial x_q}|_{(\tau_{pp}^i, X^{p-1}, U^{p-1}+\rho_{pp}^i(U^p - U^{p-1})} \frac{\partial x_q^{p-1}}{\partial u_r^j} \Big) \Bigg) \\
&- \sum_{l=1}^{n} \Big( \frac{\partial x_l^{p-1}}{\partial u_r^j} b_{il}^p + x_l^{p-1} \frac{\partial b_{il}^p}{\partial {u_r}^j} \Big),
\end{aligned}$$

where

$$\psi(\rho_{pk}^i) = \begin{cases} \rho_{pk}^i, & k = j, \\ 1 - \rho_{pk}^i, & k = j+1, \\ 0, & k > j+1. \end{cases}$$

**DISTRIBUTION A. Approved for public release: distribution unlimited.**

Using Algorithm A, we propose the following gradient-based search algorithm for solving (3.18)–(3.19):

**Algorithm B**

1. For a given positive integer $N$, let $t_j = jh$ for $j = 0, 1, \ldots, N$, where $h = T/N$.
2. Set $k = 0$. Choose a tolerance $\varepsilon > 0$ and an initial value $U^{(0)} = (U^0, U^1, ..., U^N)^{(0)} \in \mathbb{R}^{m\times(N+1)}$.
3. Calculate $X^{(k)} = (X^0, X^1, ..., X^N)^{(k)} \in \mathbb{R}^{n\times(N+1)}$ using (3.19).
4. Use Algorithm A and $X^{(k)}, U^{(k)}$ obtained to calculate $\left(\frac{\partial B_k^p}{\partial u_r^j}\right)^{(k)}$ and $\left(\frac{\partial C_k^p}{\partial u_r^j}\right)^{(k)}$ for $p = 1, 2, \ldots, N$, $j = 0, 1, 2, \ldots, N$, and $r = 1, 2, \ldots, m$.
5. Solve (4.4) for $\left(\frac{\partial X^p}{\partial u_r^j}\right)^{(k)} = \left(\frac{\partial x_1^p}{\partial u_r^j}, \frac{\partial x_2^p}{\partial u_r^j}, \ldots, \frac{\partial x_n^p)}{\partial u_r^j}\right)^{(k)\top}$.
6. Compute $\nabla F_h(X^{(k)}, U^{(k)}) = \left(\frac{\partial F_h}{\partial U^0}, \frac{\partial F_h}{\partial U^1}, \ldots, \frac{\partial F_h}{\partial U^N}\right)^{(k)\top}$ using (4.2) and (4.3), where $\frac{\partial F_h}{\partial U^j} = \left(\frac{\partial F_h}{\partial u_1^j}, \frac{\partial F_h}{\partial u_2^j}, \ldots, \frac{\partial F_h}{\partial u_m^j}\right)^\top$ for $j = 0, 1, \ldots, N$. If $||\nabla F_h(X^{(k)}, U^{(k)})|| < \varepsilon$, goto Step 8. Otherwise, continue.
7. Compute $(X^{(k+1)}, U^{(k+1)})$ using the backtracking line search method as follows:
   - 7a. Choose $\sigma^0 > 0$, $\beta \in (0, 1)$ and $\gamma \in (0, 1)$. Let $l = 1$.
   - 7b. Update $\sigma^l = \beta\sigma^{l-1}$.
   - 7c. Let $\hat{U} = U^{(k)} - \sigma^l \nabla F_h(X^{(k)}, U^{(k)})$ and compute $\hat{X}$ using (3.19) and $\hat{U}$.
   - 7d. If $F_h(\hat{X}, \hat{U}) \le F_h(X^{(k)}, U^{(k)}) - \gamma\sigma^l ||\nabla F_h(X^{(k)}, U^{(k)})||^2$, set $U^{(k+1)} = \hat{U}$, $X^{(k+1)} = \hat{X}$ and $k = k + 1$, and goto Step 4. Otherwise, let $l = l + 1$ and goto Step 7b.
8. Let $(X^*, U^*) = (X^{(k)}, U^{(k)})$ and evaluate $F_h(X^*, U^*)$ using (3.18).

**Remark:** When using Algorithm B, we need to solve the two systems (3.14) and (4.4). Note that the square coefficient matrices in these two systems are the same. Thus, we only need to calculate one matrix inverse. It is also worth pointing out that when implementing Algorithm A we can calculate $(\frac{\partial B^p}{\partial u_r^j}, \frac{\partial C^p}{\partial u_r^j})$ for $t_j, j = 1, 2, \ldots, N$ in parallel. This can save a considerable amount of CPU time when $N$ is large .

# 5 Numerical Results

In this section, we will use Algorithm B to solve several non-trivial examples. In our numerical experiments, we choose $\beta = 0.6$, $\gamma = 0.05$ and $\sigma_0 = 10$ in Algorithm B.

**DISTRIBUTION A. Approved for public release: distribution unlimited.**

**Example 1.** Consider the following FOCP with two state variables

$$\begin{aligned} \min \quad & F(u) = \frac{1}{2}\int_0^1 \left(x_1^2(t) + x_2^2(t) + u^2(t)\right)dt, \\ \text{subject to} \quad & \begin{cases} {}_0D_t^{\alpha_1} x_1(t) = -x_1(t) + x_2(t) + u(t), \\ {}_0D_t^{\alpha_2} x_2(t) = -2x_2(t), \\ x_1(0) = x_2(0) = 1. \end{cases} \end{aligned} \tag{5.1}$$

The exact solution for $\alpha_1 = \alpha_2 = 1$ is

$$x_1(t) = -\frac{3}{2}e^{-2t} + 2.48164e^{-\sqrt{2}t} + 0.018352e^{\sqrt{2}t}, \tag{5.2}$$

$$x_2(t) = e^{-2t}, \tag{5.3}$$

$$u(t) = \frac{1}{2}e^{-2t} - 1.02793e^{-\sqrt{2}t} + 0.0443056e^{\sqrt{2}t}. \tag{5.4}$$

This test problem is taken from [7]. Substituting (5.2), (5.3) and (5.4) into (5.1), we find the exact cost $F$ is

$$F = 0.431984$$

We first solve the problem for various choices of $N$ when $\alpha_1 = \alpha_2 = 1$. The optimal values of $F$ are listed in Table (5.1). It is clear from Table 5.1 that the computed optimal

Table 5.1: Optimal cost $F$ for different choices of $N$ with $\alpha_1 = \alpha_2 = 1$ for Example 1.

| $N$ | 20 | 40 | 80 | 160 | 320 |
|---|---|---|---|---|---|
| $F$ | 0.432743 | 0.432180 | 0.432036 | 0.431999 | 0.431990 |

cost approaches the theoretical one as $N$ becomes larger and all the computed costs over-estimate the exact one.

We then solve Example 1 for various choices of $\alpha_1$ and $\alpha_2$ when $N = 200$. The results are shown in Table 5.2 and Figures 5.1–5.3. From Table 5.2 we see that the total cost decreases as the values of $\alpha_1$ and $\alpha_2$ decrease.

Table 5.2: Optimal value of $F$ for different choices of $\alpha_1, \alpha_2$ for Example 1; $N = 200$

| $\alpha$ | $\alpha_1 = \alpha_2 = 1$ | $\alpha_1 = 0.9, \alpha_2 = 0.5$ | $\alpha_1 = 0.2, \alpha_2 = 0.3$ |
|---|---|---|---|
| $F$ | 0.431995 | 0.347784 | 0.267591 |

**Example 2.** Consider the following FOCP with bound constraints on the control.

$$\begin{aligned} \min \quad & F(u) = \frac{1}{2}\int_0^1 \left(x_1^2(t) + x_2^2(t) + u^2(t)\right) dt, \\ \text{subject to} \quad & \begin{cases} {}_0D_t^{\alpha_1} x_1(t) = -x_1(t) + x_2(t) + u(t), \\ {}_0D_t^{\alpha_2} x_2(t) = -2x_2(t), \\ x_1(0) = x_2(0) = 1, \end{cases} \\ & -0.2 \le u(t) \le 0. \end{aligned}$$

**DISTRIBUTION A. Approved for public release: distribution unlimited.**

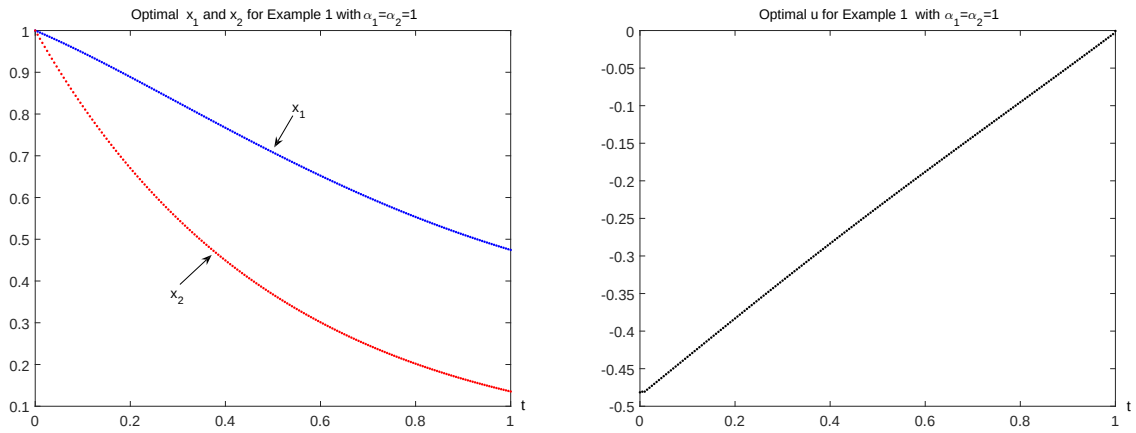



Figure 5.1: Optimal values of $x_1, x_2$ and $u$ for Example 1 with $\alpha_1 = \alpha_2 = 1$

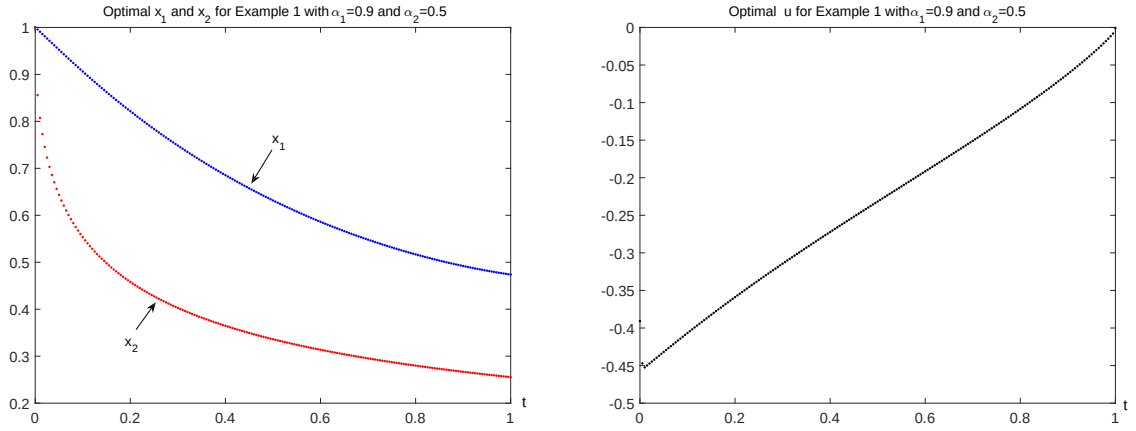



Figure 5.2: Optimal values of $x_1, x_2$ and $u$ for Example 1 with $\alpha_1 = 0.9, \alpha_2 = 0.5$

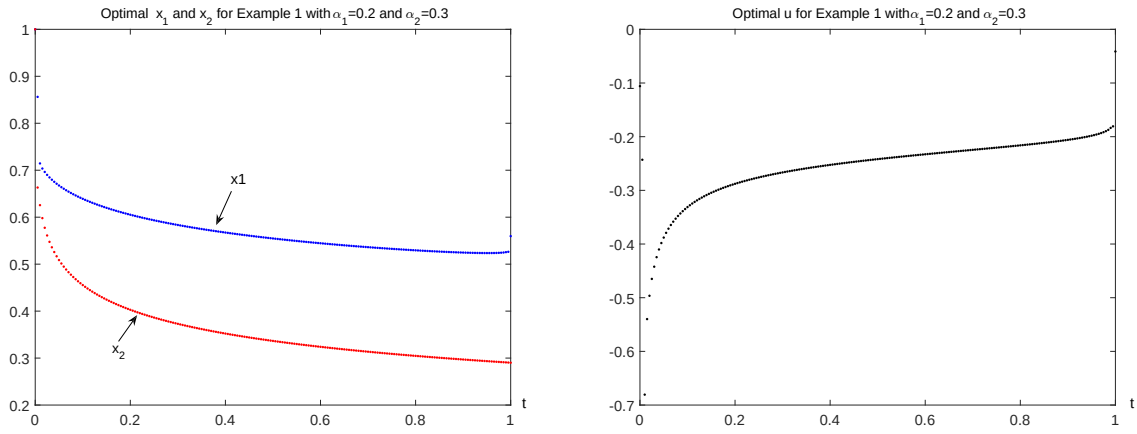



Figure 5.3: Optimal values of $x_1, x_2$ and $u$ for Example 1 with $\alpha_1 = 0.2, \alpha_2 = 0.3$

This example is the same as Example 1 with lower and upper bounds on the control $u$.

To solve this problem, we first transform the FOCP to the following form:

$$\min \ F(u) = \frac{1}{2}\int_0^1 \left(x_1^2(t) + x_2^2(t) + u^2(t)\right) dt + \lambda \int_0^1 \left((u^2(t)_+ + (-u(t) - 0.2)_+^2\right) dt,$$

subject to the system of dynamic constraints

$$\begin{aligned} {}_0D_t^{\alpha_1} x_1(t) &= -x_1(t) + x_2(t) + u(t), \\ {}_0D_t^{\alpha_2} x_2(t) &= -2x_2(t), \\ x_1(0) &= x_2(0) = 1. \end{aligned}$$

We choose $\lambda = 10,000$. The computed optimal values of $F$ corresponding to different values of $\alpha$ are shown in Table 5.3 in which we also list the computed optimal costs of Example 1 for comparison. From the table we see that computed optimal costs for the constrained problem are slightly bigger than the corresponding optimal costs of the unconstrained problem which is reasonable. To further see the difference between results from the unconstrained and constrained problems, we plot the optimal values of $x_1(t), x_2(t)$ and $u(t)$ for the different values of $\alpha$ in Figures 5.5–5.6. From the figures we see that $u$ satisfies the constraints.

Table 5.3: Optimal value of $F$ from different choice of $\alpha_1, \alpha_2$ for Example 1 and 2; $N = 200$

| - | $\alpha_1 = \alpha_2 = 1$ | $\alpha_1 = 0.9, \alpha_2 = 0.5$ | $\alpha_1 = 0.2, \alpha_2 = 0.3$ |
|---|---|---|---|
| Example 1 | 0.431995 | 0.347784 | 0.267591 |
| Example 2 | 0.442711 | 0.356319 | 0.272743 |

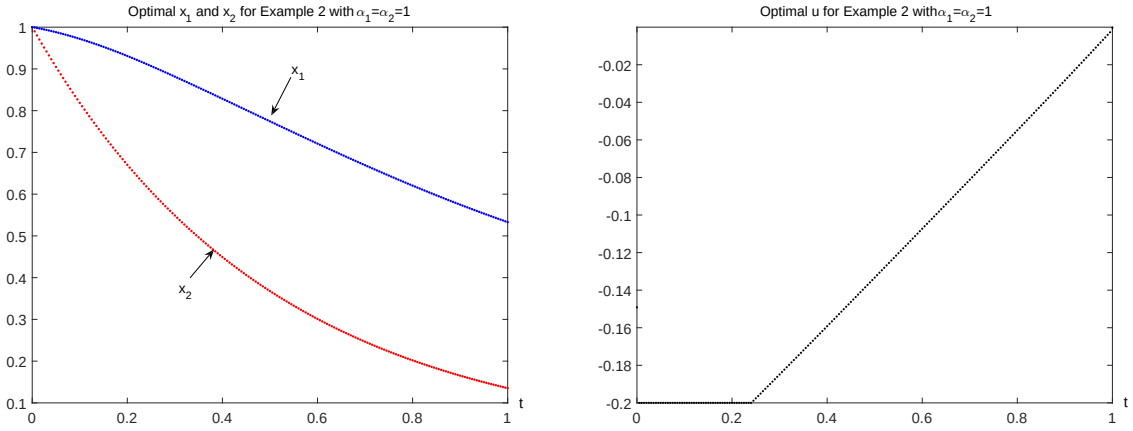



Figure 5.4: Optimal values of $x_1, x_2$ and $u$ for Example 2 with $\alpha_1 = \alpha_2 = 1$

**Example 3.** Consider the following FOCP containing two states and two controls.

$$\min \quad F(u) = \frac{1}{2}\int_0^1 \left(x_1^2(t) + x_2^2(t) + u_1^2(t) + u_2^2(t)\right) dt,$$

$$\text{subject to} \quad \begin{cases} {}_0D_t^{\alpha_1} x_1(t) = -x_1(t) + x_2(t) + u_1(t), \\ {}_0D_t^{\alpha_2} x_2(t) = -2x_2(t) + u_2(t), \\ x_1(0) = x_2(0) = 1. \end{cases}$$

DISTRIBUTION A. Approved for public release: distribution unlimited.

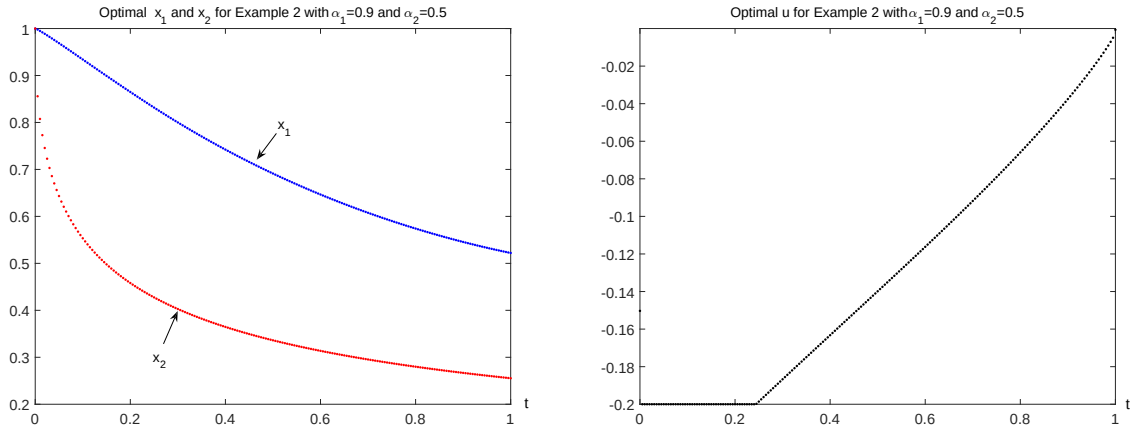



Figure 5.5: Optimal values of $x_1, x_2$ and $u$ for Example 2 with $\alpha_1 = 0.9, \alpha_2 = 0.5$

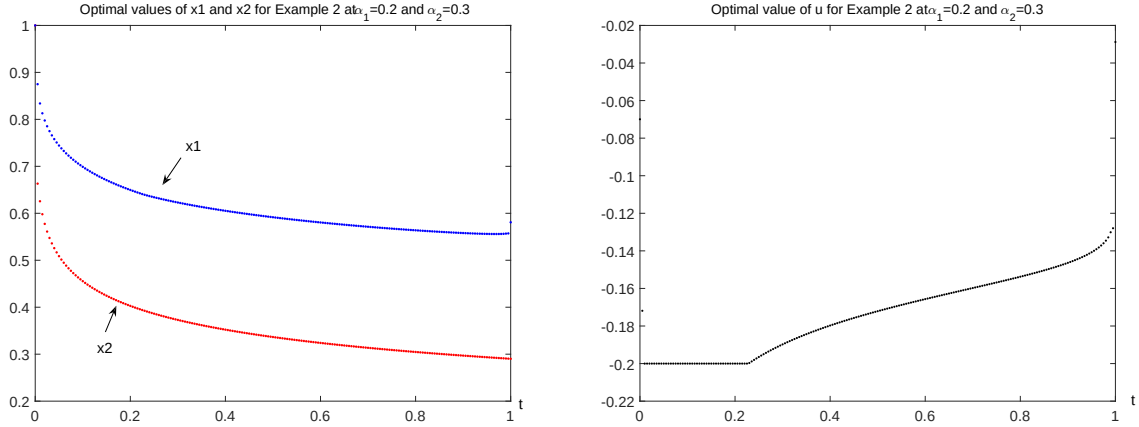



Figure 5.6: Optimal values of $x_1, x_2$ and $u$ for Example 2 with $\alpha_1 = 0.2, \alpha_2 = 0.3$

This problem is an extension of Example 1 and reduces to Example 1 when $u_2 = 0$. The optimal values of $F$ corresponding to various values of $\alpha$ are listed in Table 5.4 in which we also list the optimal costs from Example 1 for comparison. From the table we see that optimal costs from Example 3 are smaller than the corresponding ones from Example 1. This is expected as Example 1 is a special case of Example 3. We also plot some of the optimal states and controls in Figures 5.7-5.9.

Table 5.4: Optimal value of F at different choice of $\alpha_1, \alpha_2$ for Example 1 and 3 with $N = 200$

| - | $\alpha_1 = \alpha_2 = 1$ | $\alpha_1 = 0.9, \alpha_2 = 0.5$ | $\alpha_1 = 0.2, \alpha_2 = 0.3$ |
|---|---|---|---|
| Example 1 | 0.431995 | 0.347784 | 0.267591 |
| Example 3 | 0.417228 | 0.332577 | 0.259506 |

**DISTRIBUTION A. Approved for public release: distribution unlimited.**

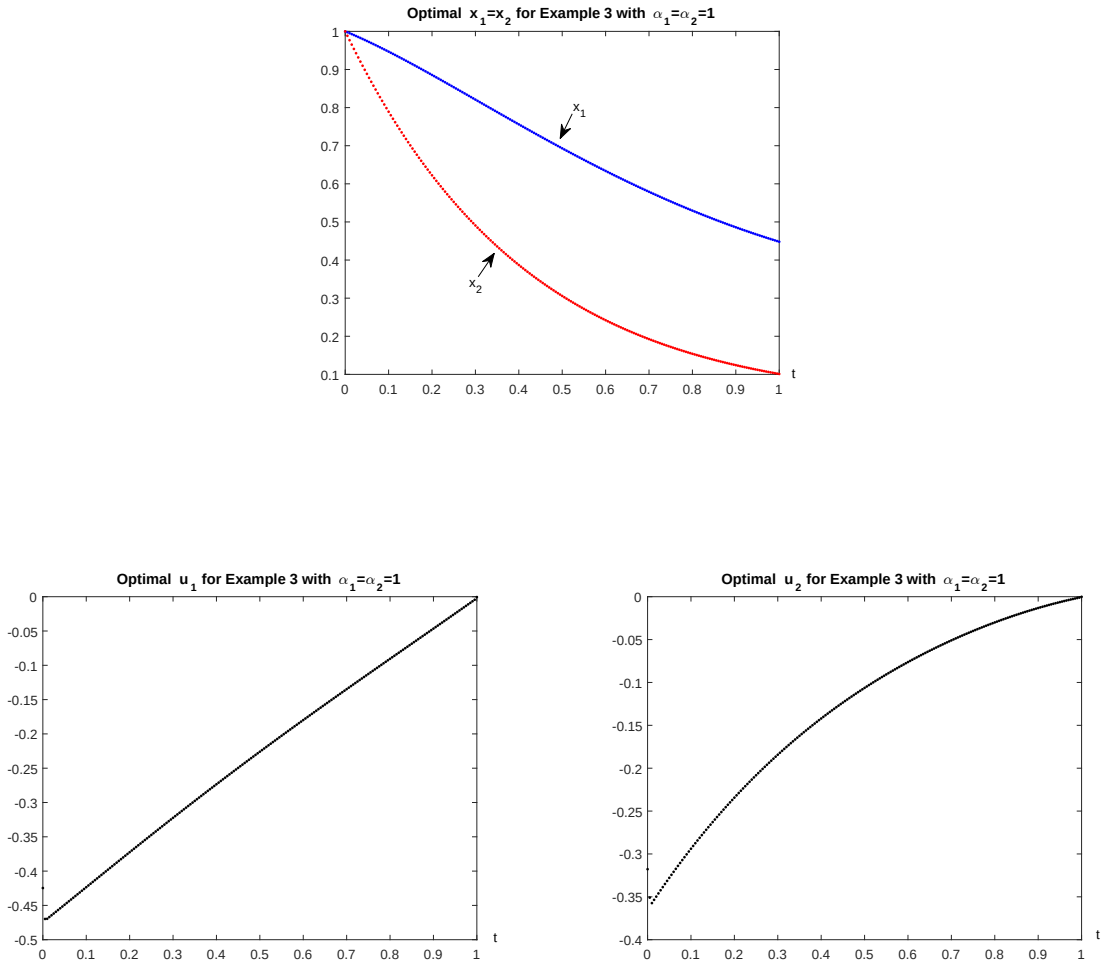



Figure 5.7: Optimal values of $x_1$, $x_2$ and $u_1, u_2$ for Example 3 with $\alpha_1 = \alpha_2 = 1$

**Example 4.**

$$\begin{aligned} \min \quad & F(u) = \int_0^{10} \left(x_1^2(t) + x_2^2(t) + u^2(t)\right) dt + x_1^2(10) + x_2^2(10), \\ \text{subject to} \quad & \begin{cases} {}_0D_t^{\alpha_1} x_1(t) = x_2(t), \\ {}_0D_t^{\alpha_2} x_2(t) = 0.2x_2(t) - x_1(t) - 0.1x_1^3(t) + u(t), \\ x_1(0) = 2, \quad x_2(0) = 0. \end{cases} \end{aligned}$$

When $\alpha = (1, 1)$, the dynamical system in this example is the Duffing equation which is known to display chaotic behaviour without any controls. We first solve the problem for various choices of $N$ when $\alpha_1 = \alpha_2 = 1$ to demonstrate that our method converges. The computed optimal values of $F$ are listed in Table 5.5. As can be seen from the table, the numerical solution improves as $N$ increases. The computed $x_1, x_2$ and $u$ when $N = 1000$ and $\alpha_1 = \alpha_2 = 1$ are shown in Figure 5.10.

**DISTRIBUTION A. Approved for public release: distribution unlimited.**

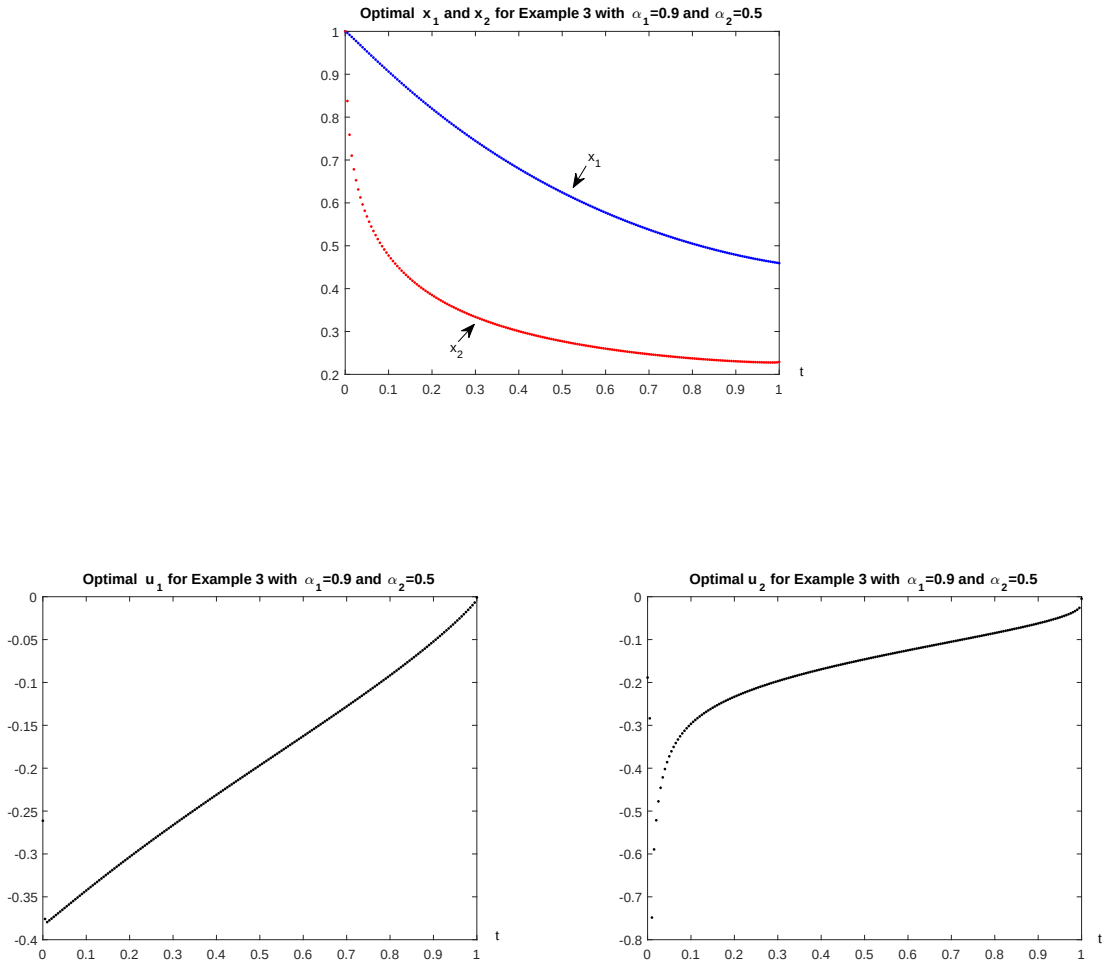



Figure 5.8: Optimal values of $x_1$, $x_2$ and $u_1, u_2$ for Example 3 with $\alpha_1 = 0.9, \alpha_2 = 0.5$

| $N$ | 300 | 600 | 800 | 1000 | 1200 |
|---|---|---|---|---|---|
| $F$ | 10.5906 | 9.3284 | 9.2114 | 9.1659 | 9.1464 |

Table 5.5: Optimal costs of Example 4 when $\alpha_1 = \alpha_2 = 1$.

We now solve Example 4 when $\alpha_1 = 0.7, \alpha_2 = 0.8$ using the uniform mesh with $N = 1200$. The computed optimal value of $F$ is 6.9390 and the computed optimal $x_1, x_2$ and $u$ are shown in Figures 5.11. From this example we see that the fractional optimal control can achieve a better optimal solution than its integer counterpart. In fact, this phenomenon is true for all of our examples in this section.

**DISTRIBUTION A. Approved for public release: distribution unlimited.**

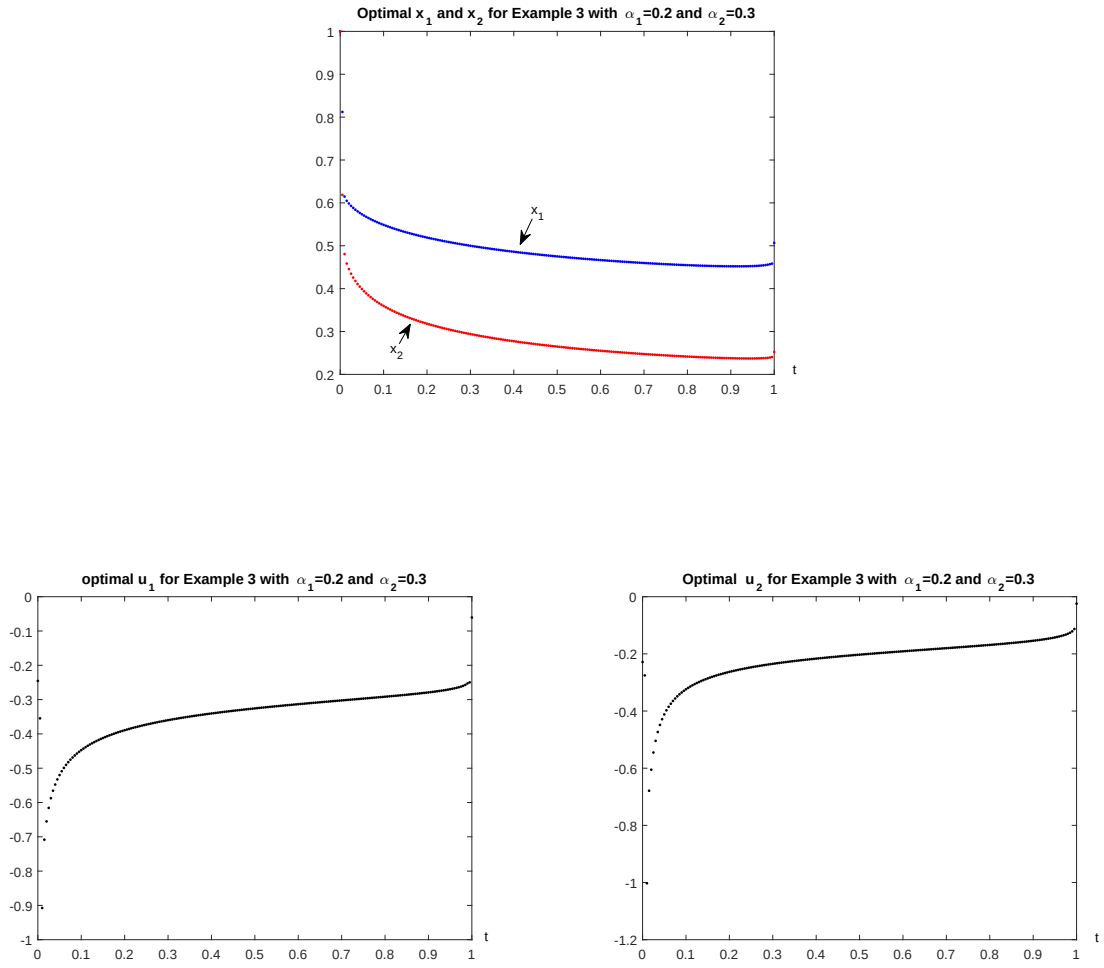



Figure 5.9: Optimal $x_1$, $x_2$ and $u_1, u_2$ for Example 3 with $\alpha_1 = 0.2, \alpha_2 = 0.3$

# 6 Conclusion

In this paper, we presented a numerical method for solving nonlinear fractional optimal control problems with multiple states and controls. We first devised a novel 2nd-order numerical integration technique for the fractional dynamical system using a set of judiciously chosen quadrature points. Based on this numerical integration technique, we then proposed a scheme for the discretization of the continuous fractional optimal control problem. A formula for calculating the gradient of the discretized cost function with respect to the decision variables has been derived and a gradient-based algorithm has been proposed for finding an optimal solution to the discretized optimal control problem. Numerical experiments on several non-trivial fractional optimal control problems have been conducted and numerical results from these experiments show that our method is accurate and robust with respect to the orders of the fractional systems. The numerical results also show that the method is able to solve real-world fractional optimal control

**DISTRIBUTION A. Approved for public release: distribution unlimited.**

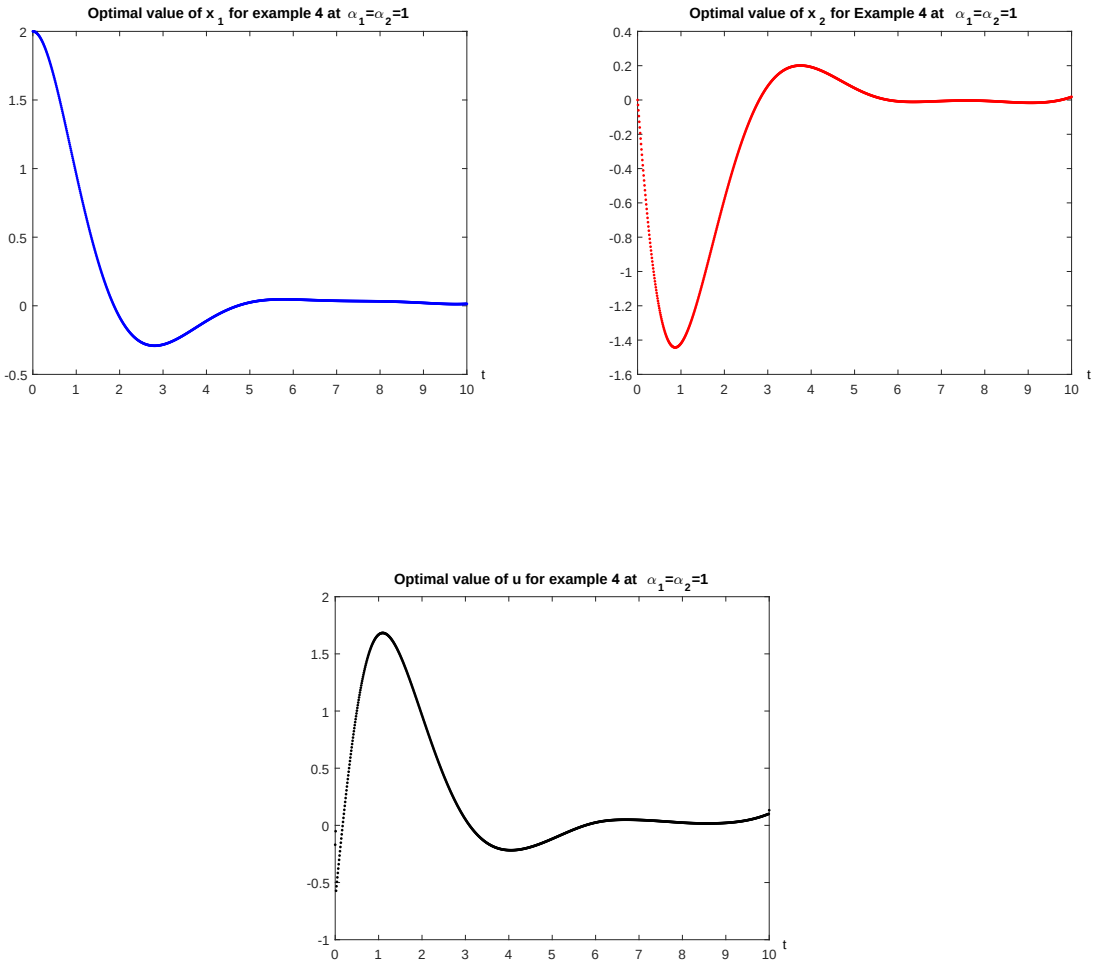



Figure 5.10: Optimal values of $x_1$, $x_2$ and $u$ for Example 4 with $\alpha_1 = \alpha_2 = 1$

problems with multiple states and controls.

# References


[1] Agrawal O.P., A general formulation and solution scheme for fractional optimal control problems, Nonlinear Dynamics, 38, 323–337, 2004.

[2] Agrawal O.P., A Hamiltonian formulation and a direct numerical scheme for fractional optimal control problems, J. Vib. Control, 13, 1269–1281, 2007.

[3] Agrawal O.P., A quadratic numerical scheme for fractional optimal control problems, J. Dyn. Sys. Meas. Control, 130, 011010, 2008.


**DISTRIBUTION A. Approved for public release: distribution unlimited.**

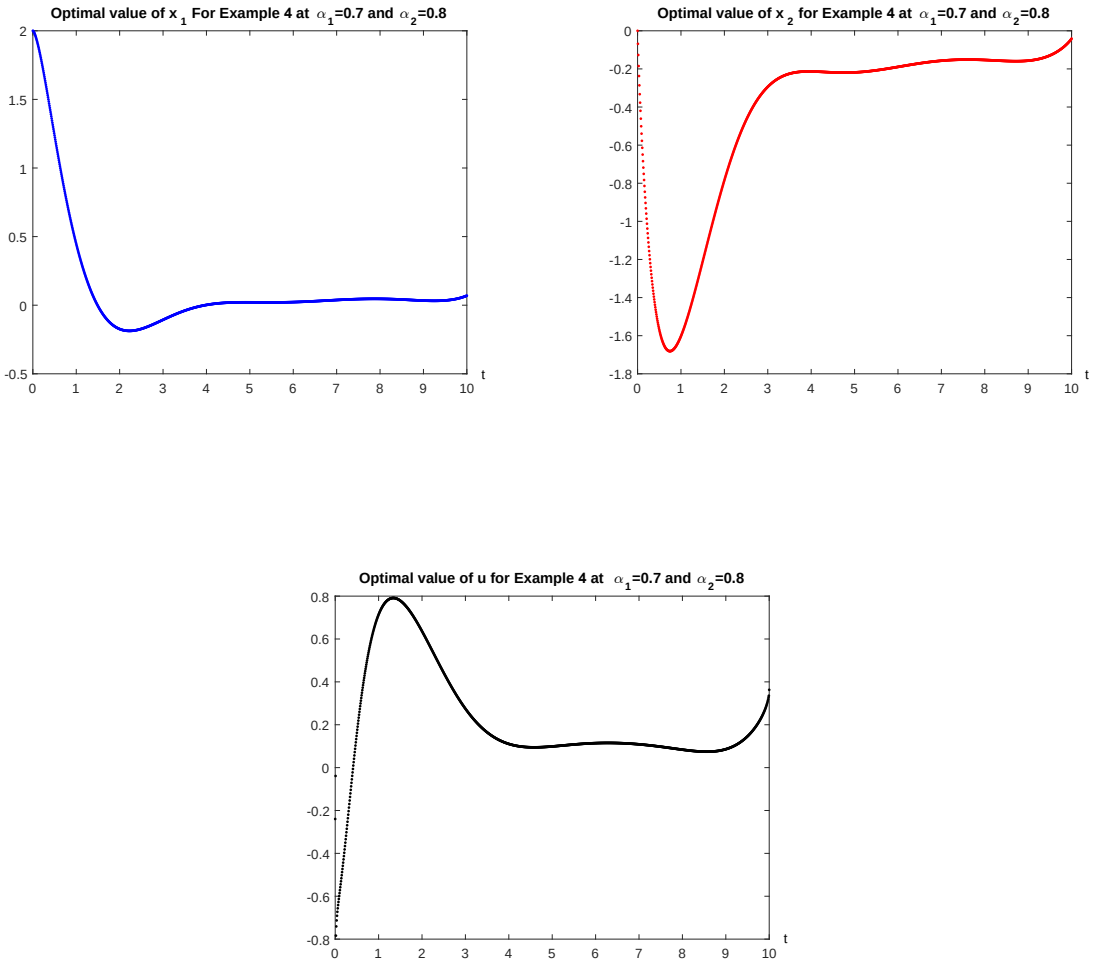



Figure 5.11: Optimal values of $x_1$ and $x_2$ for Example 4 with $\alpha_1 = 0.7$ and $\alpha_2 = 0.8$

[4] Alipour M., Rostamy D., Baleanu D., Solving multi-dimensional fractional optimal control problems with inequality constraint by Bernstein polynomials operational matrices, J. Vib. Control 19, 2523–2540, 2013.

[5] Alizadeh A, Effati S, An iterative approach for solving fractional optimal control problems, J. Vib. Control 1–19, 2016. doi:10.1177/1077546316633391, .

[6] Baleanu, D., Defterli, O., Agrawal, O.P., A central difference numerical scheme for fractional optimal control problems, J. Vib. Control, 15, 583–597, 2009

[7] Bhrawy A.H., Doha E.H., MAchado J.A., Ezz-Eldien S.S., An efficient numerical scheme for solving multi-dimensional fractional optimal control problems with a quadratic performance index, Asian Journal of Control, 17, 2389-2402, 2015.

[8] Deng W. and Li C., Numerical schemes for fractional ordinary differential equations, Numerical Modeling, Dr. Peep Miidla (Ed.), ISBN: 978-953-51-0219-9, InTech.

[9] Cao J. and Xu C., A high order scheme for the numerical solution of the fractional ordinary differential equations, Journal of Computational Physics, 238, 154–168, 2013.

[10] Chen W. and Wang S., A penalty method for a fractional order parabolic variational inequality governing American put option valuation Computers & Mathematics with Applications, 67, 77–90, 2014.

[11] Chen W. and Wang S., A power penalty method for a 2D fractional partial differential linear complementarity problem governing two-asset American option pricing Applied MAthematics & Computation, 305, 174–187, 2017.

[12] Dehghan M., Hamedi E.A., Khosravian-Arab H., A numerical scheme for the solution of a class of fractional variational and optimal control problems using the modified Jacobi polynomials, Journal of Vibration and Control, 22(6): 1547–1559, 2016.

[13] Di Pillo G. and Grippo L. Exact penalty functions in constrained optimization, SIAM J. Control and Optimization, 27, 1333–1360, 1989.

[14] Diethelm, K. and Ford, N.J.,*Analysis of fractional differential equations*, Journal of Mathematical Analysis and Applications, 265, 229–248, 2002.

[15] Diethelm K. Ford N.J. and Freed A.D., A predictor corrector approach for the numerical solution of fractional differential equation, Nonlinear Dynam, 29, No. 1-4, 2–22, 2002.

[16] Diethelm K. Ford, N.J. and Freed A.D., Detailed error analysis for a fractional Adams method, Numer. Algorithms, 36, No.1, 31–52, 2004.

[17] Diethelm K., Ford N.J. Free, A.D. and Yu L., Algorithms for the fractional calculus: a selection of numerical methods, Comput. Method Appl. Mech. Engrg., 194, 743–773, 2005.

**DISTRIBUTION A. Approved for public release: distribution unlimited.**

[18] Doha E.H., Bhrawy A.H., Baleanu D., Ezz-Eldien S.S., Hafez R.M., An efficient numerical scheme based on the shifted orthonormal Jacobi polynomials for solving fractional optimal control problems, Adv. Differ. Equ., 2015, 2015:15. doi:10.1186/s13662-014-0344-z.

[19] Ezz-Eldien S.S., Doha E.H., Baleanu D., Bhrawy A.H., A numerical approach based on Legendre orthonormal polynomials for numerical solutions of fractional optimal control problems, Journal of Vibration and Control, 23, 16–30, 2017.

[20] Glowinski G., *Numerical Methods for Nonlinear Variational Problems* Springer-Verlag, New York-Berlin-Heidelberg-Tokyo, 1984.

[21] Guo B., Pu X. and Huang F., Fractional partial differential equations and their numerical solutions, World Scientific, 2015.

[22] Huang H., Tang Y. and Vazquez L., Convergence analysis of a block-by block method for fractional differential equation. Numer. Math. Theor. Methods Appl. 5, 229–241, 2012.

[23] Kilbas A.A. and Marzan, S.A., Cauchy problem for differential equation with Caputo derivative, Fractional Calculus and Applied Analysis, 7, 288–321, 2004.

[24] Kumar K. and Agrawal O.P., An approximate method for numerical solution of fractional differential equations, Signal Process, 86, 2602–2610, 2006.

[25] Lotfi A., Dehghan M., Yousefi SA. A numerical technique for solving fractional optimal control problems, Comput. Math. Appl., 62, Issue 3, 1055–1067, 2011.

[26] Lotfi A., Yousefi SA., Dehghan M., Numerical solution of a class of fractional optimal control problems via the Legendre orthonormal basis combined with the operational matrix and the Gauss quadrature rule, Journal of Computational and Applied Mathematics, 250, 143–160, 2013.

[27] Li C. and Tao C., On the fractional Adams method, Comput. Math. Appl., 58, 1573–1588, 2009.

[28] Li, C. and Zeng, F., The finite difference methods for fractional ordinary differential equations, Numerical Functional Analysis and Optimization, 34, 149–179, 2013.

[29] Li W. and Wang S., Pricing American options under proportional transaction costs using a penalty approach and a finite difference scheme. *Journal of Industrial and Management Optimization*, 9, 365–389, 2013.

[30] Li, W., Wang, S. and Rehbock, V., A 2nd-order One-step Numerical Integration Scheme for a Fractional Differential Equation, Numerical Algebra, Control and Optimization, 7, 273–287, 2017.

[31] Lin, R. and Liu. F., Fractional high order methods for the nonlinear fractional ordinary differential equation, Nonlinear Analysis, 66, 856–869, 2007.

**DISTRIBUTION A. Approved for public release: distribution unlimited.**

[32] Nemati A., Yousefi S., Soltanian F., Ardabili J.S., An efficient numerical solution of fractional optimal control problems by using the Ritz method and Bernstein operational matrix,Asian Journal of Control, 18, 2272–2282, 2016.

[33] Nocedal J., Wright S.J. *Numerical Optimization* Springer, New York, 1999.

[34] Podlubny, L., Fractional differential equations, Mathematics in Science an Engineering, 198, Academic Press, San Diego, 1999.

[35] Rubinov A.M. and Yang X.Q. *Lagrange-type Functions in Constrained Non-Convex Optimization*. Kluwer Academic Publishers, 2003.

[36] Singha N, Nahak C, An Efficient Approximation Technique for Solving a Class of Fractional Optimal Control Problems, Journal of Optimization Theory and Application, 174, 785–802, 2017.

[37] Tricaud C, Chen YQ, An approximation method for numerically solving fractional order optimal control problems of general form, Journal of Computers and Mathematics with Applications, 59, 1644-1655, 2010.

[38] Xing A.Q., Chen Z.H., Wang C.L., Yao Y.Y., Exact penalty function approach to constrained optimal control problems, Optimal Control Applications & Methods, 10, 173–180, 1989.

**DISTRIBUTION A. Approved for public release: distribution unlimited.**

NUMERICAL ALGEBRA,
CONTROL AND OPTIMIZATION
Volume **7**, Number **3**, September **2017**

doi:10.3934/naco.2017018

pp. **273–287**

# A 2ND-ORDER ONE-POINT NUMERICAL INTEGRATION SCHEME FOR FRACTIONAL ORDINARY DIFFERENTIAL EQUATIONS

Wen Li, Song Wang and Volker Rehbock

Department of Mathematics & Statistics
Curtin University, GPO Box U1987, Perth WA 6845, Australia

(Communicated by Kok Lay Teo)

Abstract. In this paper we propose an efficient and easy-to-implement numerical method for an $\alpha$-th order Ordinary Differential Equation (ODE) when $\alpha \in (0, 1)$, based on a one-point quadrature rule. The quadrature point in each sub-interval of a given partition with mesh size $h$ is chosen judiciously so that the degree of accuracy of the quadrature rule is 2 in the presence of the singular integral kernel. The resulting time-stepping method can be regarded as the counterpart for fractional ODEs of the well-known mid-point method for 1st-order ODEs. We show that the global error in a numerical solution generated by this method is of the order $\mathcal{O}(h^2)$, independently of $\alpha$. Numerical results are presented to demonstrate that the computed rates of convergence match the theoretical one very well and that our method is much more accurate than a well-known one-step method when $\alpha$ is small.

1. **Introduction.** Modelling and optimal control of many practical systems in engineering, science and economics traditionally involve Ordinary Differential Equation (ODE) systems of integer orders [2, 24, 25, 27, 28, 29, 30]. While integer order ODE systems are adequate for capturing the evolution of most standard phenomena, it has been shown over the last two decades that many complex systems in solid mechanics, viscoelastics, gas diffusion and heat conduction in porous media, signal and image processing, bio-engineering, biology, economics and financial engineering are better modelled by systems with fractional or non-integer-order differential equations (cf., for example, [3, 4, 5, 6, 7, 8, 22, 23, 26]). In particular, complex phenomena involving memory effects can be modelled more appropriately and accurately by fractional dynamical systems than by classical (integer) ones. As it is very rare that a system of practical significance can be solved analytically, one needs to be able to solve the system numerically. Clearly, an accurate and computationally efficient numerical scheme is crucial for solving fractional ODEs. This is particularly true when solving an optimal control problem involving such a system as an iterative computational procedure for computing the optimal control requires the solving the system repeatedly.

2010 *Mathematics Subject Classification.* Primary: 65L05, 65L20; Secondary: 49M25.

*Key words and phrases.* Fractional ordinary differential equation, optimal control, one-step time-stepping scheme, one-point quadrature rule, super-convergence, Caputo's fractional derivative.

This work is supported by the AOARD Project # 15IOA095 from the US Air Force.

* Corresponding author: S. Wang.



**DISTRIBUTION A. Approved for public release: distribution unlimited.**

We consider the following fractional initial value problem:

$$ {}_0D_t^\alpha x(t) = f(t, x(t)), \quad t \in (0, T], \tag{1} $$

$$ x(0) = x_0, \tag{2} $$

where $x_0$ and $T$ are positive constants. $f$ is a known function and ${}_0D_t^\alpha x(t)$ denotes the following Captuto's $\alpha$th-order derivative of $x(t)$ in $(0, T]$ with $\alpha \in (0, 1)$:

$$ {}_0D_t^\alpha x(t) = \frac{1}{\Gamma(1-\alpha)} \int_0^t \frac{x'(\tau)}{(t-\tau)^\alpha}\, d\tau. $$

In the above $\Gamma(\cdot)$ denotes the Gamma function. Higher-order fractional initial value problems can be transformed into a system of fractional initial value problems of the form (1)-(2) and any efficient numerical method developed for (1)–(2) can be extended to a vector-valued initial value problem. There is also another representation of the $\alpha$th-derivative called the Riemann-Liouville fractional derivative. However, initial value problems involving the Riemann-Liouville fractional derivative can be readily transformed into (1)–(2) as demonstrated in [10, 14]. It has been proven [10, 14, 16] that solving (1)-(2) is equivalent to solving the following Volterra integral equation:

$$ x(t) = x_0 + \frac{1}{\Gamma(\alpha)} \int_0^t (t-\tau)^{\alpha-1} f(\tau, x(\tau))\, d\tau, \quad \alpha \in (0, 1). \tag{3} $$

In the open literature, there are four main numerical methods for solving (3): Euler's method [20], an Adams type predictor-corrector method [11, 12, 13, 19, 20], the $p$-th order method [20, 21] and the block by block method [1, 15, 17]. Euler's method is simple, but the convergence order is only $\mathcal{O}(h^\alpha)$, where $h$ denotes the mesh size of a uniform partition of $(0, T)$. The Adams type predictor-corrector method was first proposed by Diethelem et al. [11]. They showed that the convergence order of this method is $\mathcal{O}(h^{1+\alpha})$. Based on the work of Diethelem et al., Deng and Li [9] have developed another improved predictor-corrector method. They proved that the order of convergence of the improved version is $\mathcal{O}(h^{\min(1+2\alpha,2)})$. Both of the schemes in [9] and [11] are single step methods. The $p$-th order and block-by-block methods have convergence rates of orders $\mathcal{O}(h^3)$ and $\mathcal{O}(h^{3+\alpha})$, respectively. However, these methods are linear multiple step methods and thus computationally much more expensive than single step methods. Therefore, a question that arises is whether it is possible to design a single step method for (3) with an upper error bound better than $\mathcal{O}(h^{\min(1+2\alpha,2)})$ when $\alpha < 0.5$. In the integer case that $\alpha = 1$, the mid-point one step method has an upper error bound of order $\mathcal{O}(h^2)$. This method takes advantage of the property that the mid-point quadrature rule yields a 'super-convergence' point for numerical integration, i.e., the mid-point is the only one in an interval which gives the exact numerical integral when the integrand is a linear polynomial. Clearly, this super-convergence property of the mid-point quadrature rule does not hold true for integrals of the type (3), because the kernel becomes singular when $\tau$ approaches $t$. Thus, the question of what is the counterpart of the classical implicit mid-point numerical integration method for (3) arises.

In this work, we design a one-step numerical method for (3) which is easy to implement, computationally inexpensive, and results in a global error of order $\mathcal{O}(h^2)$ for any $\alpha \in (0, 1)$. This method can be regarded as the counterpart for fractional ODEs of the implicit mid-point numerical integration method for first-order ODEs. In this method, we choose a numerical quadrature point in each of the sub-

DISTRIBUTION A. Approved for public release: distribution unlimited.

intervals of a given partition judiciously so that the local approximation error is of a higher order than that from the conventional one-point quadrature rule. This is the counterpart of the mid-point quadrature rule in conventional numerical integration. Based on this special numerical quadrature rule, we develop a one-step numerical integration method for (3) and prove that the global error in the numerical solutions generated by this method is of order $\mathcal{O}(h^2)$.

The rest of the paper is organized as the follows. In Section 2, we propose an approximation of (3) based on a Taylor expansion. An error analysis of the approximation is also presented. In Section 3, we propose an algorithm for implementing the approximate equation and analyse its convergence. In Section 4 we apply our method to several factional ODEs with known exact solutions to confirm the theoretical error estimate and demonstrate the superiority of our method over some of the existing ones. Section 5 concludes the paper.

2. **Approximation.** For a given positive integer $N$, we first divide $(0, T]$ into $N$ sub-intervals with mesh points $t_i = ih$ for $i = 0, 1, \ldots, N$, where $h = T/N$. Thus, (3) can be written as follows.

$$\begin{aligned} x(t_i) &= x_0 + \frac{1}{\Gamma(\alpha)} \int_0^{t_i} (t_i - \tau)^{\alpha-1} f(\tau, x(\tau))\, d\tau \\ &= x_0 + \frac{1}{\Gamma(\alpha)} \int_0^{ih} (ih - \tau)^{\alpha-1} f(\tau, x(\tau))\, d\tau \\ &= x_0 + \frac{1}{\Gamma(\alpha)} \sum_{j=1}^{i} \int_{(j-1)h}^{jh} (ih - \tau)^{\alpha-1} f(\tau, x(\tau)) d\tau. \end{aligned} \tag{4}$$

We now consider approximation of the integral on right hand side of (4). Assume that $f(t, x(t))$ is twice continuously differentiable with respect to both $t$ and $x$. For $j = 1, 2, \ldots, i$, we apply Taylor's theorem to $f(\tau, x(\tau))$ at $\tau_{ij}$ to get

$$f(\tau, x(\tau)) = f(\tau_{ij}, x(\tau_{ij})) + K_{ij}(\tau - \tau_{ij}) + c_{ij}(\tau - \tau_{ij})^2, \tag{5}$$

where $c_{ij}$ a constant representing the 2nd derivative of $f$ at a point between $\tau_{ij}$ and $\tau$ and

$$K_{ij} = f_\tau(\tau_{ij}, x(\tau_{ij})) + f_x(\tau_{ij}, x(\tau_{ij}))x'_\tau(\tau_{ij}).$$

Therefore, replacing $f(\tau, x(\tau))$ in the integrand of the last term in (4) with the RHS of (5), we have, for any $j = 1, ..., i$,

$$\begin{aligned} &\frac{1}{\Gamma(\alpha)} \int_{(j-1)h}^{jh} (ih - \tau)^{\alpha-1} f(\tau, x(\tau)) d\tau \\ &= \frac{1}{\Gamma(\alpha)} \int_{(j-1)h}^{jh} (ih - \tau)^{\alpha-1} [f(\tau_{ij}, x(\tau_{ij})) + K_{ij}(\tau - \tau_{ij})] d\tau + R_{ij} \\ &= \frac{1}{\Gamma(\alpha)} f(\tau_{ij}, x(\tau_{ij})) \left[ \frac{(ih - (j-1)h)^\alpha}{\alpha} - \frac{(ih - jh)^\alpha}{\alpha} \right] \\ &\quad + \frac{K_{ij}}{\Gamma(\alpha)} \int_{(j-1)h}^{jh} (ih - \tau)^{\alpha-1} (\tau - \tau_{ij})\, d\tau + R_{ij} \\ &= \frac{h^\alpha}{\Gamma(\alpha + 1)} f(\tau_{ij}, x(\tau_{ij}))[(i - j + 1)^\alpha - (i - j)^\alpha] \end{aligned}$$

DISTRIBUTION A. Approved for public release: distribution unlimited.

$$+\frac{K_{ij}}{\Gamma(\alpha)}\int_{(j-1)h}^{jh}(ih-\tau)^{\alpha-1}(\tau-\tau_{ij})d\tau+R_{ij}, \tag{6}$$

where $R_{ij}=\dfrac{1}{\Gamma(\alpha)}\displaystyle\int_{(j-1)h}^{jh}(ih-\tau)^{\alpha-1}c_{ij}(\tau-\tau_{ij})^2d\tau$.

We now consider the choice of $\tau_{ij}$. From (6) it is clear that $\tau_{ij}$ should be chosen such that the second term becomes zero so that the truncation error in (6) is just $R_{ij}$. The choice of $\tau_{ij}$ is given in the following theorem.

**Theorem 2.1.** *For any given $i=1,2,\dots,N$ and $j=1,2\dots,i$, the unique solution to*

$$\int_{(j-1)h}^{jh}(ih-\tau)^{\alpha-1}(\tau-\tau_{ij})d\tau=0 \tag{7}$$

*is given by*

$$\tau_{ij}=h\frac{[(i-j+1)^{\alpha+1}-(i-j)^{\alpha+1}]+(\alpha+1)[(i-j+1)^{\alpha}(j-1)-(i-j)^{\alpha}j]}{(\alpha+1)[(i-j+1)^{\alpha}-(i-j)^{\alpha}]}. \tag{8}$$

*Furthermore, $(j-1)h<\tau_{ij}<jh$.*

*Proof.* We first integrate the LHS of (7) using integration by parts as follows.

$$\begin{aligned}\int_{(j-1)h}^{jh}(ih-\tau)^{\alpha-1}(\tau-\tau_{ij})d\tau=&-\frac{(ih-\tau)^{\alpha}}{\alpha}(\tau-\tau_{ij})\Big|_{(j-1)h}^{jh}-\frac{(ih-\tau)^{\alpha+1}}{\alpha(\alpha+1)}\Big|_{(j-1)h}^{jh}\\ =&-\frac{[(i-j)^{\alpha}j-(i-j+1)^{\alpha}(j-1)]h^{\alpha+1}}{\alpha}\\ &-\frac{[(i-j+1)^{\alpha}-(i-j)^{\alpha}]h^{\alpha}}{\alpha}\tau_{ij}\\ &+\frac{(i-j+1)^{\alpha+1}-(i-j)^{\alpha+1}}{\alpha(\alpha+1)}h^{\alpha+1}\\ =&0.\end{aligned}$$

From the above, we have

$$\begin{aligned}&[(i-j+1)^{\alpha}-(i-j)^{\alpha}]\tau_{ij}\\ &=\frac{(i-j+1)^{\alpha+1}-(i-j)^{\alpha+1}}{\alpha+1}h+[(i-j+1)^{\alpha}(j-1)-(i-j)^{\alpha}j]h.\end{aligned}$$

Solving this for $\tau_{ij}$, we get (8).

We now show that $(j-1)h<\tau_{ij}<jh$. From (8),

$$\begin{aligned}&\frac{\tau_{ij}}{h}-j\\ &=\frac{[(i-j+1)^{\alpha+1}-(i-j)^{\alpha+1}]+(\alpha+1)[(i-j+1)^{\alpha}(j-1)-(i-j)^{\alpha}j]}{(\alpha+1)[(i-j+1)^{\alpha}-(i-j)^{\alpha}]}-j\\ &=\frac{[(i-j+1)^{\alpha+1}-(i-j)^{\alpha+1}]+(\alpha+1)[(i-j+1)^{\alpha}(j-1)-(i-j)^{\alpha}j]}{(\alpha+1)[(i-j+1)^{\alpha}-(i-j)^{\alpha}]}\\ &\quad-\frac{(\alpha+1)j[(i-j+1)^{\alpha}-(i-j)^{\alpha}]}{(\alpha+1)[(i-j+1)^{\alpha}-(i-j)^{\alpha}]}\end{aligned}$$

DISTRIBUTION A. Approved for public release: distribution unlimited.

$$= \frac{(i-j+1)^{\alpha+1} - (i-j)^{\alpha+1} - (\alpha+1)(i-j+1)^{\alpha}}{(\alpha+1)[(i-j+1)^{\alpha} - (i-j)^{\alpha}]}.$$

To prove $\tau_{ij} - jh < 0$, we only need to show that

$$(i-j+1)^{\alpha+1} - (i-j)^{\alpha+1} - (\alpha+1)(i-j+1)^{\alpha} < 0.$$

Using the mean value theorem, we have

$$(i-j+1)^{\alpha+1} - (i-j)^{\alpha+1} - (\alpha+1)(i-j+1)^{\alpha} = (\alpha+1)\xi^{\alpha} - (\alpha+1)(i-j+1)^{\alpha} < 0,$$

since $\xi \in (i-j, i-j+1)$. Thus, we conclude that $\tau_{ij} - jh < 0$. In a similar way, we can prove that $\tau_{ij} - (j-1)h > 0$. □

Combining Theorem 2.1 and (4), we have the following representation for $x(t_i)$.

$$x(t_i) = x_0 + \frac{h^{\alpha}}{\Gamma(\alpha+1)} \sum_{j=1}^{i} f(\tau_{ij}, x(\tau_{ij}))[(i-j+1)^{\alpha} - (i-j)^{\alpha}] + R_i \tag{9}$$

for $i = 1, 2, \ldots, N$, where $\tau_{ij}$ is given in (8) for $j = 1, 2, \ldots, i$ and $R_i = \sum_{j=1}^{i} R_{ij}$. Omitting the remainder $R_i$ in (9), we have an equation approximating (4) which has the truncation error $R_i$. An upper bound for $R_i$ is given in the following theorem.

**Theorem 2.2.** *Suppose that $f(t, x(t))$ is twice continuously differentiable in $t$ and $x$. Then, for any $i = 1, 2, .., N$, the following estimate holds:*

$$|R_i| \le Ch^2, \tag{10}$$

*where $C$ denotes a positive constant independent of $h$ and $\alpha$.*

*Proof.* For $j = 1, 2, \ldots, i$, from the definition of $R_i$ and Theorem 2.1 we have

$$\begin{aligned} |R_{ij}| &= \left| \frac{c_{ij}}{\Gamma(\alpha)} \int_{(j-1)h}^{jh} (ih-\tau)^{\alpha-1} (\tau - \tau_{ij})^2 d\tau \right| \\ &\le \frac{|c_{ij}|h^2}{\Gamma(\alpha)} \int_{(j-1)h}^{jh} (ih-\tau)^{\alpha-1} d\tau \\ &= \frac{|c_{ij}|h^2}{\Gamma(\alpha)} \left\{ \frac{[(i-j+1)h]^{\alpha}}{\alpha} - \frac{[(i-j)h]^{\alpha}}{\alpha} \right\}. \end{aligned}$$

Since $f$ is twice continuously differentiable, $c_{ij}$ is bounded on $[0, T]$. Let $c = \max_i(\max_{1 \le j \le i} |c_{ij}|)$. Then, from the definition of $R_i$ and the above estimate, we have

$$\begin{aligned} |R_i| &\le \sum_{j=1}^{i} |R_{ij}| \\ &\le \frac{ch^2}{\Gamma(\alpha)} \sum_{j=1}^{i} \left\{ \frac{[(i-j+1)h]^{\alpha}}{\alpha} - \frac{[(i-j)h]^{\alpha}}{\alpha} \right\} \\ &= \frac{ch^2}{\Gamma(\alpha)} \frac{(ih)^{\alpha}}{\alpha} \\ &\le \frac{ch^2}{\Gamma(\alpha)} \frac{T^{\alpha}}{\alpha} \\ &= C_1 h^2, \end{aligned}$$

DISTRIBUTION A. Approved for public release: distribution unlimited.

where $C_1 = \frac{c}{\Gamma(\alpha)}\frac{T^\alpha}{\alpha} = \frac{c}{\Gamma(1+\alpha)}T^\alpha$. Since $\alpha \in (0,1)$, from [18] we have $2^{\alpha-1} \le \Gamma(1+\alpha) \le 1$. Also, it is obvious that $T^\alpha < \max\{1, T\}$ for $\alpha \in (0,1)$. Therefore, from the above estimates we see that (10) holds true for a $C$ independent of both $h$ and $\alpha$. Thus, we have proved the theorem. □

From (9) it is clear that to compute $x(t_i)$, we need to calculate $f(\tau_{ij}, x(\tau_{ij})), j = 1, 2, \ldots, i$. However, $x(\tau_{ij}), j = 1, \ldots, i$, are not available directly from the scheme, although the $\tau_{ij}$ are known. Thus, approximations for $x(\tau_{ij}), j = 1, \ldots, i$, need to be determined. In the next section, we propose a single step numerical scheme for implementing (9) when the remainder $R_i$ is omitted.

3. **Algorithm and convergence.** In this section, we propose an explicit time-stepping algorithm for approximating the solution of (9) when $R_i$ is omitted, based on the linearization of the nonlinear term in $x(t_i)$ in (9).

For any indices $i$ and $j$ satisfying $1 \le j \le i \le N$, since $\tau_{ij} \in (t_{j-1}, t_j)$ by Theorem 2.1, we use the following linear interpolation to approximate $x(\tau_{ij})$:

$$x(\tau_{ij}) = x(t_{j-1}) + \rho_{ij}(x(t_j) - x(t_{j-1})) + \mathcal{O}(h^2), \tag{11}$$

where

$$\rho_{ij} := \frac{\tau_{ij} - t_{j-1}}{h} \in (0,1). \tag{12}$$

Then, we approximate $f(\tau_{ij}, x(\tau_{ij}))$ as follows.

$$f(\tau_{ij}, x(\tau_{ij})) = f\left(\tau_{ij}, x(t_{j-1}) + \rho_{ij}(x(t_j) - x(t_{j-1}))\right) + \mathcal{O}(h^2) \tag{13}$$

for $1 \le j \le i \le N$. Clearly, if we replace $f(\tau_{ij}, x(\tau_{ij}))$ in (9) with the above expression and omit the truncation error terms of order $\mathcal{O}(h^2)$, we define the following single step time-stepping scheme for (3):

$$\begin{aligned} x_i &= x_0 + h_\alpha \sum_{j=1}^{i} f\left(\tau_{ij}, x_{j-1} + \rho_{ij}(x_j - x_{j-1})\right)\left[(i-j+1)^\alpha - (i-j)^\alpha\right] \\ &= x_0 + h_\alpha g_i(x_0, x_1, ..., x_{i-1}) + h_\alpha f(\tau_{ii}, x_{i-1} + \rho_{ii}(x_i - x_{i-1})) \end{aligned} \tag{14}$$

for $i = 1, 2, \ldots, N$, where $\tau_{ij}$ is defined by (8) and

$$h_\alpha = h^\alpha/\Gamma(1+\alpha), \tag{15}$$

$$g_i(x_0, x_1, ..., x_{i-1}) = \begin{cases} 0, & i = 1, \\ \displaystyle\sum_{j=1}^{i-1} f\left(\tau_{ij}, x_{j-1} + \rho_{ij}(x_j - x_{j-1})\right)\left[(i-j+1)^\alpha - (i-j)^\alpha\right], & i > 1. \end{cases} \tag{16}$$

The above scheme is implicit as the last term on the RHS of (14) contains a nonlinear function of $x_i$. We can use an iterative method such as the conventional Newton's method to solve (14) which is usually computationally more expensive than the predictor-corrector process. However, for this case, we may define an explicit single step scheme by further approximating the last term in (14) by the following truncated Taylor's expansion:

$$f(\tau_{ii}, x_{i-1} + \rho_{ii}(x_i - x_{i-1})) = f(\tau_{ii}, x_{i-1}) + f_x(\tau_{ii}, x_{i-1})\rho_{ii}(x_i - x_{i-1}) + \mathcal{O}(h^2). \tag{17}$$

Combining (14) and (17), we propose the following explicit one-step algorithm for (3).

DISTRIBUTION A. Approved for public release: distribution unlimited.

**Algorithm A:**

1. For a given positive integer $N$, let $t_i = ih$ for $i = 0, 1, ..., N$, where $h = T/N$.
2. For $i = 1, 2, \ldots, N$, compute

$$x_i = \frac{x_0 + h_\alpha g_i(x_0, x_1, ..., x_{i-1}) + h_\alpha\big(f(\tau_{ii}, x_{i-1}) - \rho_{ii} f_x(\tau_{ii}, x_{i-1}) x_{i-1}\big)}{1 - h_\alpha \rho_{ii} f_x(\tau_{ii}, x_{i-1})}, \tag{18}$$

where $\tau_{ij}$, $\rho_{ij}$, $h_\alpha$ and $g_i(x_0, ..., x_{i-1})$ are defined in (8), (12), (15) and (16), respectively.

As can be seen later, Algorithm A provides an efficient and stable 2nd-order method for (3). Strictly speaking, Algorithm A is a multiple step method. This is because the fractional derivative is a global operator and thus all $x_j, j < i$ are needed in order to evaluate $x_i$ in Algorithm A. However, since the last two terms in the numerator on the RHS of (18) only involve $x_{i-1}$, we still call it a one-step method. Also, unlike the case of the explicit mid-point method for first-order ODEs, it is generally very hard to construct a one-step explicit or predictor-correct method based on (9). This is because construction of such a scheme usually requires a fractional Taylor expansion which typically has a truncation error of order $\mathcal{O}(h^{n\alpha})$ for some positive integer $n$. Clearly, $n\alpha \to 0$ as $\alpha$ approaches 0.

Using linear interpolation theory and Taylor's theorem, we are able to prove that, for any $i = 1, 2, \ldots, N$, $x_i$ generated by Algorithm A converges to $x(t_i)$ at the rate $\mathcal{O}(h^2)$ when $h \to 0^+$. This is formalized in the following theorem.

**Theorem 3.1.** *Let $x(t)$ be the exact solution of (3)/(4). If $f(t, x)$ is twice continuously differentiable in $t$ and $x$, then there exists an $\bar{h} > 0$ such that $h_\alpha < \bar{h}$ implies*

$$|x(t_i) - x_i| \le Ch^2, \quad i = 1, 2, \ldots, N, \tag{19}$$

*where $\{x_i\}$ is the sequence generated by Algorithm A, $C$ is a positive constant, independent of $h$ and $\alpha$, and $h_\alpha$ is defined in (15).*

*Proof.* In what follows, we let $C$ denote a generic positive constant, independent of $h$. We now prove this theorem by mathematical induction. Firstly, we show that (19) holds for $i = 1$.

By (9), we have

$$x(t_1) = x_0 + h_\alpha f(\tau_{11}, x(\tau_{11})) + R_1.$$

Furthermore, from (11) and (13), we have

$$\begin{aligned} x(t_1) &= x_0 + h_\alpha f(\tau_{11}, x_0 + \rho_{11}(x(t_1) - x_0) + \mathcal{O}(h^2) + R_1 \\ &= x_0 + h_\alpha \{f(\tau_{11}, x_0) + f_x(\tau_{11}, x_0)[\rho_{11}(x(t_1) - x_0) + \mathcal{O}(h^2)] + \mathcal{O}(h^2)\} + R_1 \\ &= x_0 + h_\alpha \left[f(\tau_{11}, x_0) - f_x(\tau_{11}, x_0)\rho_{11} x_0\right] + h_\alpha f_x(\tau_{11}, x_0)\rho_{11} x(t_1) + \mathcal{O}(h^2) \\ &\quad + R_1, \end{aligned}$$

where $\rho_{11}$ is defined in (12). Solving this equation for $x(t_1)$ and using (18) with $i = 1$ we have

$$\begin{aligned} x(t_1) &= \frac{x_0 + h_\alpha \left[f(\tau_{11}, x_0) - f_x(\tau_{11}, x_0)\rho_{11} x_0\right] + \mathcal{O}(h^2) + R_1}{1 - h_\alpha f_x(\tau_{11}, x_0)\rho_{11}} \\ &= x_1 + \frac{\mathcal{O}(h^2) + R_1}{1 - h_\alpha f_x(\tau_{11}, x_0)\rho_{11}}. \end{aligned}$$

DISTRIBUTION A. Approved for public release: distribution unlimited.

Therefore, using (10), the previous equation yields the inequality

$$|x(t_1) - x_1| \le \frac{Ch^2}{|1 - h_\alpha f_x(\tau_{11}, x_0)\rho_{11}|}. \tag{20}$$

Note $\rho_{11} \in (0,1)$ by (12) and $h_\alpha > 0$. So, if $f_x(\tau_{11}, x_0) \le 0$, from (20) we see that (19) is satisfied for $i = 1$. However, when $f_x(\tau_{11}, x_0) > 0$, we need to choose an upper bound for $h_\alpha$ so that the denominator of (20) is bounded below by a positive constant. Clearly, for a given constant $\sigma \in (0,1)$, if we choose $\bar{h}_1 := \frac{1-\sigma}{\max\{\rho_{11} f_x(\tau_{11}, x_0), 1\}}$, then (19) is satisfied for $i = 1$, when $h_\alpha \le \bar{h}_1$. To be more general, we choose

$$\bar{h} := \frac{1 - \sigma}{\max\{\max_{1 \le i \le N} \rho_{ii} f_x(\tau_{ii}, x_{i-1}), 1\}}, \tag{21}$$

and when $h_\alpha \le \bar{h}$, (19) is satisfied for $i = 1$.

We now consider the case of $i \ge 2$. Assume that

$$\max_{1 \le j \le i-1} |x(t_j) - x_j| \le Ch^2 \tag{22}$$

when $h_\alpha \le \bar{h}$. We now show that $\max_{1 \le j \le i} |x(t_j) - x_j| \le Ch^2$, or equivalently, $|x(t_i) - x_i| \le Ch^2$.

Using (13) and (17), we have, from (9) and (10)

$$\begin{aligned}
&x(t_i)\\
=&x_0 + h_\alpha \sum_{j=1}^{i} \big\{ \big[ f(\tau_{ij}, x(t_{j-1}) + \rho_{ij}(x(t_j) - x(t_{j-1})) + \mathcal{O}(h^2)) \big]\\
&\times [(i-j+1)^\alpha - (i-j)^\alpha] \big\} + R_i\\
=&x_0\\
&+ h_\alpha \sum_{j=1}^{i-1} f(\tau_{ij}, x(t_{j-1}) + \rho_{ij}(x(t_j) - x(t_{j-1})) + \mathcal{O}(h^2))[(i-j+1)^\alpha - (i-j)^\alpha]\\
&+ h_\alpha \left[ f(\tau_{ii}, x(t_{i-1})) + f_x(\tau_{ii}, x(t_{i-1}))\rho_{ii}(x(t_i) - x(t_{i-1})) \right] + \mathcal{O}(h^2) + R_i\\
=&x_0 + A_{i-1} + B_i + \mathcal{O}(h^2).
\end{aligned} \tag{23}$$

where

$$\begin{aligned}
A_{i-1} =& h_\alpha \sum_{j=1}^{i-1} f(\tau_{ij}, x(t_{j-1}) + \rho_{ij}(x(t_j) - x(t_{j-1}))\\
&+ \mathcal{O}(h^2))[(i-j+1)^\alpha - (i-j)^\alpha],
\end{aligned} \tag{24}$$

$$B_i = h_\alpha [f(\tau_{ii}, x(t_{i-1})) + f_x(\tau_{ii}, x(t_{i-1}))\rho_{ii}(x(t_i) - x(t_{i-1}))]. \tag{25}$$

Note that (18) can be re-written in the following form.

$$\begin{aligned}
x_i &= x_0 + h_\alpha \sum_{j=1}^{i-1} f(\tau_{ij}, x_{j-1} + \rho_{ij}(x_j - x_{j-1}))[(i-j+1)^\alpha - (i-j)^\alpha]\\
&\quad + h_\alpha [f(\tau_{ii}, x_{i-1}) + f_x(\tau_{ii}, x_{i-1})\rho_{ii}(x_i - x_{i-1})]\\
&= x_0 + \tilde{A}_{i-1} + \tilde{B}_i,
\end{aligned} \tag{26}$$

DISTRIBUTION A. Approved for public release: distribution unlimited.

where

$$\tilde{A}_{i-1} = h_\alpha \sum_{j=1}^{i-1} f(\tau_{ij}, x_{j-1} + \rho_{ij}(x_j - x_{j-1}))[(i-j+1)^\alpha - (i-j)^\alpha], \quad (27)$$

$$\tilde{B}_i = h_\alpha \left[ f(\tau_{ii}, x_{i-1}) + f_x(\tau_{ii}, x_{i-1})\rho_{ii}(x_i - x_{i-1}) \right]. \quad (28)$$

Subtracting (26) from (23) gives

$$x(t_i) - x_i = (A_{i-1} - \tilde{A}_{i-1}) + (B_i - \tilde{B}_i) + \mathcal{O}(h^2). \quad (29)$$

Let us first estimate $B_i - \tilde{B}_i$. From (25) and (28), we have

$$\begin{aligned} B_i - \tilde{B}_i =& [h_\alpha [f(\tau_{ii}, x(t_{i-1})) + f_x(\tau_{ii}, x(t_{i-1}))\rho_{ii}(x(t_i) - x(t_{i-1}))] \\ & - h_\alpha [f(\tau_{ii}, x_{i-1}) + f_x(\tau_{ii}, x_{i-1})\rho_{ii}(x_i - x_{i-1})]] \\ =& h_\alpha [f(\tau_{ii}, x(t_{i-1})) - f(\tau_{ii}, x_{i-1})] \\ & + h_\alpha \rho_{ii} [f_x(\tau_{ii}, x(t_{i-1}))x(t_i) - f_x(\tau_{ii}, x_{i-1})x_i] \\ & - h_\alpha \rho_{ii} [f_x(\tau_{ii}, x(t_{i-1}))x(t_{i-1}) - f_x(\tau_{ii}, x_{i-1})x_{i-1}]. \end{aligned} \quad (30)$$

Note that $f$ is twice continuously differentiable. Using a Taylor expansion, we get

$$f_x(\tau_{ii}, x(t_{i-1})) = f_x(\tau_{ii}, x_{i-1}) + r_i, \quad (31)$$

where

$$r_i = f_{xx}(\tau_{ii}, \xi)(x(t_{i-1}) - x_{i-1})$$

with $\xi$ being a point between $x(t_{i-1})$ and $x_{i-1}$. Since $|x(t_{i-1}) - x_{i-1}| \le Ch^2$ by Assumption (22), we have

$$r_i = f_{xx}(\tau_{ii}, \xi) \cdot \mathcal{O}(h^2) = \mathcal{O}(h^2).$$

Similarly, we have

$$f(\tau_{ii}, x(t_{i-1})) - f(\tau_{ii}, x_{i-1}) = \mathcal{O}(h^2).$$

Using (31) and the above estimate we have, from (30),

$$\begin{aligned} B_i - \tilde{B}_i =& h_\alpha \mathcal{O}(h^2) + h_\alpha \rho_{ii} [f_x(\tau_{ii}, x_{i-1})(x(t_i) - x_i) + x(t_i) r_i] \\ & - h_\alpha \rho_{ii} [f_x(\tau_{ii}, x_{i-1})(x(t_{i-1})) - x_{i-1}) + x(t_{i-1}) r_i] \\ =& h_\alpha \rho_{ii} f_x(\tau_{ii}, x_{i-1})[x(t_i) - x_i] + h_\alpha \rho_{ii} f_x(\tau_{ii}, x_{i-1})[x_{i-1} - x(t_{i-1})] \\ & + \mathcal{O}(h^{2+\alpha}) \\ =& h_\alpha \rho_{ii} f_x(\tau_{ii}, x_{i-1})[x(t_i) - x_i] + \mathcal{O}(h^{2+\alpha}), \end{aligned}$$

since $h_\alpha = h^\alpha / \Gamma(1+\alpha)$ and $|x_{i-1} - x(t_{i-1})| \le Ch^2$ from Assumption (22). Thus, from the above expression and (29), we get

$$x(t_i) - x_i = (A_{i-1} - \tilde{A}_{i-1}) + h_\alpha \rho_{ii} f_x(\tau_{ii}, x_{i-1})[x(t_i) - (x_i)] + \mathcal{O}(h^2).$$

This implies

$$x(t_i) - x_i = \frac{(A_{i-1} - \tilde{A}_{i-1}) + \mathcal{O}(h^2)}{1 - h_\alpha \rho_{ii} f_x(\tau_{ii}, x_{i-1})},$$

and so

$$|x(t_i) - x_i| \le \frac{|A_{i-1} - \tilde{A}_{i-1}|}{|1 - h_\alpha \rho_{ii} f_x(\tau_{ii}, x_{i-1})|} + \frac{\mathcal{O}(h^2)}{|1 - h_\alpha \rho_{ii} f_x(\tau_{ii}, x_{i-1})|} \quad (32)$$

To estimate $|x(t_i) - x_i|$, we need to estimate $|A_{i-1} - \tilde{A}_{i-1}|$. To simplify our notation, we let $x_{ij} = x_{j-1} + \rho_{ij}(x_j - x_{j-1})$. We also use either the RHS or LHS

DISTRIBUTION A. Approved for public release: distribution unlimited.

of (11) to represent the point $x(\tau_{ij})$. From the definitions of $A_{i-1}$ and $\tilde{A}_{i-1}$ in (24) and (27), respectively, and using (11), we have

$$\begin{aligned}|A_{i-1}-\tilde{A}_{i-1}| =& h_\alpha \left|\sum_{j=1}^{i-1}[f(\tau_{ij},x(\tau_{ij}))-f(\tau_{ij},x_{ij})][(i-j+1)^\alpha-(i-j)^\alpha]\right| \\ \le& h_\alpha \sum_{j=1}^{i-1}|[f(\tau_{ij},x(\tau_{ij}))-f(\tau_{ij},x_{ij})][(i-j+1)^\alpha-(i-j)^\alpha]| \\ =& h_\alpha \sum_{j=1}^{i-1}|f(\tau_{ij},x(\tau_{ij}))-f(\tau_{ij},x_{ij})|\,[(i-j+1)^\alpha-(i-j)^\alpha], \quad (33)\end{aligned}$$

since $z^\alpha$ is an increasing function of $z$ for $\alpha\in(0,1)$. Because $f$ is twice continuously differentiable, we have (recall that $C$ is a generic positive constant, independent of $h$)

$$\begin{aligned}|f(\tau_{ij},x(\tau_{ij}))-f(\tau_{ij},x_{ij})| \le& C|x(\tau_{ij})-x_{ij}| \\ =& C\Big|[x(t_{j-1})+\rho_{ij}(x(t_j)-x(t_{j-1}))+\mathcal{O}(h^2)] \\ & -[x_{j-1}+\rho_{ij}(x_j-x_{j-1})]\Big| \\ =& C\Big|[x(t_{j-1})-x_{j-1}]+\rho_{ij}[x(t_j)-x_j)] \\ & +\rho_{ij}(x_{j-1}-x(t_{j-1}))]\Big|+\mathcal{O}(h^2) \\ \le& C\left(|x(t_{j-1})-x_{j-1}|+|x(t_j)-x_j)|+|x(t_{j-1})-x_{j-1}|\right) \\ & +\mathcal{O}(h^2),\end{aligned}$$

since $\rho_{ij}\in(0,1)$. In the above we have used (11). Thus, from Assumption (22), we have

$$|f(\tau_{ij},x(\tau_{ij}))-f(\tau_{ij},x_{\tau_{ij}})|\le Ch^2.$$

Replacing $|f(\tau_{ij},x(\tau_{ij}))-f(\tau_{ij},x_{\tau_{ij}})|$ in (33) with the above upper bound, we have

$$\begin{aligned}|A_{i-1}-\tilde{A}_{i-1}| \;\le\; & h_\alpha Ch^2\sum_{j=1}^{i-1}[(i-j+1)^\alpha-(i-j)^\alpha] \\ =& \frac{h^\alpha}{\Gamma(\alpha+1)}Ch^2(i^\alpha-1) \\ \le& \frac{h^\alpha}{\Gamma(\alpha+1)}Ch^2N^\alpha \\ =& \frac{C}{\Gamma(\alpha+1)}h^2(hN)^\alpha \\ =& \frac{C}{\Gamma(\alpha+1)}h^2T^\alpha \\ \le& Ch^2.\end{aligned}$$

Combining the above error bound with (32), we have

$$|x(t_i)-x_i|\le\frac{Ch^2}{|1-h_\alpha\rho_{ii}f_x(\tau_{ii},x_{i-1})|}.$$

DISTRIBUTION A. Approved for public release: distribution unlimited.

Therefore, when $h_\alpha \le \bar{h}$, where $\bar{h}$ is defined in (21), we have

$$|x(t_i) - x_i| \le \frac{Ch^2}{|1 - h_\alpha \rho_{ii} f_x(\tau_{ii}, x_{i-1})|} \le \frac{Ch^2}{\sigma}$$

for a given $\sigma > 0$.

Careful readers may have noticed the positive constant $C$ used in the above proof is independent of $h$, but a function of $T^\alpha/\Gamma(1+\alpha)$. However, in the proof of Theorem 2.2 we have shown that $T^\alpha/\Gamma(1+\alpha)$ is bounded above by a positive constant, independent of $\alpha$. Thus, we have proved the theorem. □

We comment that in the above theorem we have established an upper error bound of order $\mathcal{O}(h^2)$ for the numerical approximation to (1)–(2) generated by our proposed single-step Algorithm A, while all of the existing single-step methods proposed in [9, 11, 12, 13, 19, 20] have the drawback that their rates of convergence approach $\mathcal{O}(h)$ as $\alpha$ decreases.

We also comment that Algorithm A is a linearized form of our implicit method (14). This linearization does not affect the $\mathcal{O}(h^2)$-order rate of convergence of (14). In other words, our explicit method represented in Algorithm A performs only one Newton iteration for (14) as performing more Newton iterations will not increase the accuracy of the numerical method due to the discretization errors.

## 4. Numerical Results.

In this section, we solve two examples using our proposed method.

**Example 1.** Consider the following fractional differential equation

$$\begin{aligned} {}_0D_t^\alpha x(t) &= \frac{\Gamma(5)}{\Gamma(5-\alpha)} t^{4-\alpha} - x(t) + t^4, \quad t \in (0,1], \\ x(0) &= 0. \end{aligned}$$

The exact solution is

$$x(t) = t^4.$$

This test problem is taken from [9] and it is solved by Algorithm A for various values of $\alpha$ and mesh sizes $h$. The computed errors $E_{h_k} = \max_{1 \le i \le 1/h_k} |x_i - x(t_i)|$ for $h_k = 1/(2^k \times 10), k = 0, 1, ..., 6$ and the chosen values of $\alpha$ are listed in Table 1. To estimate the rates of convergence, we calculate $\log_2(E_{h_k}/E_{h_{k+1}})$ for $k = 0, 1, ..., 5$ and the computed rates of convergence are also listed in Table 1. From the table we see that the computed rates of convergence are very close to the theoretical one in Theorem 3.1. In Table 1, we also compare our results with those obtained by the method in [9]. The latter method is a single-step predictor-corrector method with a theoretical rate of convergence of order $\mathcal{O}(h^{\min(1+2\alpha,2)})$. The rates of convergence of our method are much higher than those of the method in [9] for $\alpha = 0.1$. From the table we also see that the absolute errors from our method are much smaller than those in [9] unless $\alpha$ is close to 1 in which case classic methods apply. Clearly, our proposed method is superior to that in [9]. In addition, one can expect that a predictor-corrector method should be computationally more expensive than our method.

From Table 1 we see that when $\alpha$ is close to zero, the computed rate of convergence is slight worse than $\mathcal{O}(h^2)$. This may be because the constant $C$ in (19) contains the term $1/\Gamma(1+\alpha)$, as noted in the previous section. However, when $h$ decreases, the rate of convergence of our method increases.

DISTRIBUTION A. Approved for public release: distribution unlimited.

TABLE 1. Maximum Errors and Convergence Rates for Example 1.

| $h$ | Our results | | Results from [9] | | Our results | | Results from [9] | |
|---|---|---|---|---|---|---|---|---|
| | $\alpha$=0.1 | Order | $\alpha$=0.1 | Order | $\alpha$=0.3 | Order | $\alpha$=0.3 | Order |
| 1/10 | 4.06e-3 | - | 0.364 | - | 7.67e-3 | - | - | - |
| 1/20 | 1.11e-3 | 1.86 | 0.170 | 1.10 | 2.00e-3 | 1.94 | - | - |
| 1/40 | 3.02e-4 | 1.89 | 7.13e-2 | 1.26 | 5.17 e-4 | 1.95 | - | - |
| 1/80 | 8.08e-5 | 1.90 | 2.88e-2 | 1.31 | 1.32e-4 | 1.97 | - | - |
| 1/160 | 2.14e-5 | 1.92 | 1.15e-2 | 1.32 | 3.37e-5 | 1.97 | - | - |
| 1/320 | 5.65e-6 | 1.93 | 4.64e-3 | 1.31 | 8.55e-6 | 1.98 | - | - |
| 1/640 | 1.47e-6 | 1.93 | 1.88e-3 | 1.30 | 2.16e-6 | 1.98 | - | - |
| 1/1280 | 3.84e-7 | 1.94 | - | - | 5.46e-7 | 1.99 | - | - |

| $h$ | Our results | | Results from [9] | | Our results | | Results from [9] | |
|---|---|---|---|---|---|---|---|---|
| | $\alpha$=0.5 | Order | $\alpha$=0.5 | Order | $\alpha$=0.9 | Order | $\alpha$=0.9 | Order |
| 1/10 | 8.49e-3 | - | 0.0355 | - | 7.88e-3 | - | 0.0107 | - |
| 1/20 | 2.16e-3 | 1.98 | 0.00879 | 2.01 | 1.97e-3 | 2.00 | 0.00231 | 2.21 |
| 1/40 | 5.43e-4 | 1.99 | 2.16e-3 | 2.03 | 4.93e-4 | 2.00 | 5.21e-4 | 2.15 |
| 1/80 | 1.37e-4 | 1.99 | 5.31e-4 | 2.02 | 1.23e-4 | 2.00 | 1.22e-4 | 2.09 |
| 1/160 | 3.43e-5 | 2.00 | 1.31e-4 | 2.02 | 3.08e-5 | 2.00 | 2.94e-5 | 2.06 |
| 1/320 | 8.58e-6 | 2.00 | 3.24e-5 | 2.02 | 7.70e-6 | 2.00 | 7.18e-6 | 2.03 |
| 1/640 | 2.15e-6 | 2.00 | 8.03e-6 | 2.01 | 1.92e-6 | 2.00 | 1.77e-6 | 2.01 |
| 1/1280 | 5.38e-7 | 2.00 | - | - | 4.81e-7 | 2.00 | - | - |

**Example 2.** Consider the following fractional differential equation

$$\begin{aligned} {}_0D_t^\alpha x(t) &= \frac{\Gamma(4+\alpha)}{6}t^3 + t^{3+\alpha} - x(t), \quad t \in (0,1], \\ x(0) &= 0. \end{aligned}$$

The exact solution is $x(t) = t^{3+\alpha}$. This example is from [1] and it is solved using Algorithm A for various values of $h$ and $\alpha$. The computed errors and rates of convergence are listed in Table 2 from which we see that the computed rates of convergence are close to the theoretical one in Theorem 3.1.

TABLE 2. Maximum Errors and Convergence Rates for Example 2.

| $h$ | $\alpha$=0.1 | Order | $\alpha$=0.3 | Order | $\alpha$=0.5 | Order | $\alpha$=0.9 | Order |
|---|---|---|---|---|---|---|---|---|
| 1/10 | 2.37e-3 | - | 5.08e-3 | - | 6.40e-3 | - | 7.50e-3 | - |
| 1/20 | 6.45e-4 | 1.88 | 1.31e-3 | 1.95 | 1.62e-3 | 1.98 | 1.88e-3 | 2.00 |
| 1/40 | 1.73e-4 | 1.90 | 3.39e-4 | 1.96 | 4.08e-4 | 1.99 | 4.69e-4 | 2.00 |
| 1/80 | 4.60e-5 | 1.91 | 8.65e-5 | 1.97 | 1.03e-4 | 1.99 | 1.17e-4 | 2.00 |
| 1/160 | 1.21e-5 | 1.92 | 2.20e-5 | 1.98 | 2.57e-5 | 2.00 | 2.93e-5 | 2.00 |
| 1/320 | 3.18e-6 | 1.93 | 5.58e-6 | 1.98 | 6.44e-6 | 2.00 | 7.32e-6 | 2.00 |
| 1/640 | 8.30e-7 | 1.94 | 1.41e-6 | 1.98 | 1.61e-6 | 2.00 | 1.83e-6 | 2.00 |
| 1/1280 | 2.15e-7 | 1.95 | 3.55e-7 | 1.99 | 4.04e-7 | 2.00 | 4.57e-7 | 2.00 |

**Example 3.** Consider the following fractional differential equation

$$ {}_0D_t^\alpha x(t) = \frac{\Gamma(4+\alpha)}{6}t^3 + t^{4(3+\alpha)} - x^4(t), \quad t \in (0,1], $$

DISTRIBUTION A. Approved for public release: distribution unlimited.

$$x(0) \quad = \quad 0.$$

The exact solution is also $x(t) = t^{3+\alpha}$. Note that the RHS of the above equation is nonlinear in both $t$ and $x$. This example is solved using Algorithm A for various values of $h$ and $\alpha$ and the computed errors and rates of convergence are listed in Table 3. From the table we see that the computed order of convergence is greater than 2 when $\alpha$ is small.

TABLE 3. Maximum Errors and Convergence Rates for Example 3.

| $h$ | $\alpha$=0.1 | Order | $\alpha$=0.3 | Order | $\alpha$=0.5 | Order | $\alpha$=0.9 | Order |
|---|---|---|---|---|---|---|---|---|
| 1/10 | 6.19e-2 | - | 4.27e-2 | - | 2.56e-2 | - | 3.70e-3 | - |
| 1/20 | 1.69e-2 | 1.87 | 1.03e-2 | 2.05 | 5.70e-3 | 2.17 | 6.90e-4 | 2.42 |
| 1/40 | 4.42e-3 | 1.93 | 2.368e-3 | 2.13 | 1.10e-3 | 2.37 | 1.88e-4 | 1.88 |
| 1/80 | 1.10e-3 | 2.00 | 5.055e-4 | 2.23 | 1.85e-4 | 2.57 | 5.92e-5 | 1.67 |
| 1/160 | 2.65e-4 | 2.05 | 1.025e-4 | 2.30 | 2.63e-5 | 2.82 | 2.11e-5 | 1.49 |
| 1/320 | 6.26e-5 | 2.08 | 2.00e-5 | 2.36 | 2.63e-6 | 3.32 | 6.19e-6 | 1.77 |
| 1/640 | 1.46e-5 | 2.10 | 3.77e-6 | 2.41 | 5.30e-7 | 2.31 | 1.68e-6 | 1.88 |
| 1/1280 | 3.41e-6 | 2.10 | 6.86e-7 | 2.46 | 1.55e-7 | 1.77 | 4.37e-7 | 1.94 |

To summarise, from Tables 1–3 we see that, although computed convergence rates fluctuate for different values of $\alpha$ and $h$, when $h^\alpha$ is small enough, the convergence order is 2 confirming our theoretical analysis. This can also be seen from the last row of each of the tables corresponding to $h = 1/1280$. All the errors are of the magnitude $h^2 \approx 6 \times 10^{-7}$. To check the robustness of our method in $\alpha$, we have also solved Example 3 for $\alpha = 10^{-i}$, $i = 1, ..., 8$ and the computed results show that the orders of convergence are roughly 2, particularly when $h$ is small. The robustness of our method may provide an effective way for solving problems with algebraic constraints, or differential algebraic equations. We will discuss this in a future paper.

5. **Conclusion.** In this paper, we proposed a new numerical method based on Taylor's theorem and linear interpolation for solving fractional differential equations. The proposed method is simple and easy to use. We have proved that the convergence order of the method is 2. The numerical results confirm our theoretical analysis.

**Acknowledgement.** This work is supported by the AOARD Project #15IOA095 from the US Air Force.

## REFERENCES

[1] J. Cao and C. Xu, A high order scheme for the numerical solution of the fractional ordinary differential equations, *Journal of Computational Physics*, **238** (2013), 154–168.
[2] S. Campbell and P. Kunkel, Solving higher index DAE optimal control problems, *Numerical Algebra, Control & Optimization*, **6** (2016), 447–472.
[3] A. Cartea and D. del-Castillo-Negrete, Fractional diffusion models of option prices in markets with jumps, *Physica A: Statistical Mechanics and its Applications*, **374** (2007), 749–763.
[4] W. Chen and S. Wang, A penalty method for a fractional order parabolic variational inequality governing American put option valuation, *Comp. Math. With Appl.*, **67** (2014), 77–90.

**DISTRIBUTION A. Approved for public release: distribution unlimited.**

[5] W. Chen and S. Wang, A finite difference method for pricing European and American options under a geometric Levy process, *Journal of Industrial & Management Optimization*, **11** (2015), 241–264.
[6] W. Chen and S. Wang, A 2nd-Order FDM for a 2D fractional Black-Scholes equation, in *Numerical Analysis and Its Applications. NAA 2016*(eds. Dimov I., Farag I., Vulkov L.), Lecture Notes in Computer Science, Springer, **10187** (2017), 46–57.
[7] W. Chen and S. Wang, A power penalty method for a 2D fractional partial differential linear complementarity problem governing two-asset American option pricing, *Applied Mathematics and Computation*, **305** (2017) 174–187.
[8] C. F. M. Coimbra, Mechanics with variable-order differential operators, *Ann. Phis. (Leipzig)*, **12** (2003), 692–703.
[9] W. Deng and C. Li, Numerical schemes for fractional ordinary differential equations, *Numerical Modeling*, Dr. Peep Miidla (Ed.), InTech, 2012.
[10] K. Diethelm and N. J. Ford, Analysis of fractional differential equations, *Journal of Mathematical Analysis and Applications*, **265** (2002), 229–248.
[11] K. Diethelm, N. J. Ford and A. D. Freed, A predictor corrector approach for the numerical solution of fractional differential equation, *Nonlinear Dynam*, **29** (2002), 2–22.
[12] K. Diethelm, N. J. Ford and A. D. Freed, Detailed error analysis for a fractional Adams method, *Numer. Algorithms*, **36** (2004), 31–52.
[13] K. Diethelm, N. J. Ford, A. D. Freed and Yu. Luchko, Algorithms for the fractional calculus: a selection of numerical methods, *Comput. Method appl. Mech. Engrg.*, **194** (2005), 743–773.
[14] B. Guo, X. Pu and F. Huang, *Fractional Partial Differential Equations and Their Numerical Solutions*, World Scientific, 2015.
[15] H. Huang, Y. Tang and L. Vazquez, Convergence analysis of a block-by block method for fractional differential equation, *Numer. Math. Theor. Methods Appl.*, **5** (2012), 229–241.
[16] A. A. Kilbas and S. A. Marzan, Cauchy problem for differential equation with Caputo derivative, *Fractional Calculus and Applied Analysis*, **7** (2014), 297–321.
[17] K. Kumar and O. P. Agrawal, An approximate method for numerical solution of fractional differential equations, *Signal Process*, **86** (2006), 2602–2610.
[18] A. Laforgia and P. Natalini, Exponential, gamma and polygamma functions: Simple proofs of classical and new inequalities, *J. Math. Anal. Appl.*, **407** (2013), 459–504.
[19] C. Li and C. Tao, On the fractional Adams method, *Comput. Math. Appl.*, **58** (2009), 1573–1588.
[20] C. Li and F. Zeng, The finite difference methods for fractional ordinary differential equations, *Numerical Functional Analysis and Optimization*, **34** (2013), 149–179.
[21] R. Lin and F. Liu, Fractional high order methods for the nonlinear fractional ordinary differential equation, *Nonlinear Analysis*, **66** (2007), 856–869.
[22] R. Magin, *Fractional Calculus in Bioengineering*, Begell House Inc., Redding, CT, 2006.
[23] F. Mainardi, *Fractional Calculus and Waves in Linear Viscoelasticity: An Introduction to Mathematical Models*, Imperial College Press, London, 2010.
[24] S. Mohd Mahali, S. Wang and X. Lou, Determination of effective diffusion coefficients of drug delivery devices by a state observer approach, *Discrete and Continuous Dynamical Systems Series B*, **17** (2011), 1119–1136.
[25] S. Mohd Mahali, S. Wang and X. Lou, Estimation of effective diffusion coefficients of drug delivery devices in a flow-through system, *Journal of Engineering Mathematics*, **87** (2014), 139–152.
[26] M. D. Ortigueria and J. A. T. Machodo, Special section: Fractional calculus applications in signals and systems, *Signal Processing*, **86** (2006), 2503–3094.
[27] B. Shen, X. Wang and C. Liu, Nonlinear state-dependent impulsive system in fed-batch culture and its optimal control, *Numerical Algebra, Control & Optimization*, **5** (2015), 369–380.

**DISTRIBUTION A. Approved for public release: distribution unlimited.**

[28] S. Sorokin and M. Staritsyn, Feedback necessary optimality conditions for a class of terminally constrained state-linear variational problems inspired by impulsive control, *Numerical Algebra, Control & Optimization*, **7** (2017), 201–210.
[29] M. Y. Tan, L. S. Jennings and S. Wang, Analysing human periodic walking at different speeds using parametrization enhancing transform in dynamic optimization, *Pacific Journal of Optimization*, **12** (2016), 557–586.
[30] Y. Wang, C. Yu and K. L. Teo, A new computational strategy for optimal control problem with a cost on changing control, *Numerical Algebra, Control & Optimization*, **6** (2016), 339–364.

Received January 2017; $1^{st}$ revision June 2017; final revision July 2017.

*E-mail address*: wen.li@curtin.edu.au
*E-mail address*: song.wang@curtin.edu.au
*E-mail address*: V.Rehbock@curtin.edu.au

**DISTRIBUTION A. Approved for public release: distribution unlimited.**

# A 2nd-Order FDM for a 2D Fractional Black-Scholes Equation

W. Chen[1] and S. Wang[2(✉)]

[1] CSIRO Data61, 34 Village Street, Docklands, VIC 3008, Australia
Wen.Chen@csiro.au
[2] Department of Mathematics and Statistics, Curtin University, GPO Box U1987, Perth 6845, Australia
Song.Wang@curtin.edu.au

**Abstract.** We develop a finite difference method (FDM) for a 2D fractional Black-Scholes equation arising in the optimal control problem of pricing European options on two assets under two independent geometric Lévy processes. We establish the convergence of the method by showing that the FDM is consistent, stable and monotone. We also show that the truncation error of the FDM is of 2nd order. Numerical experiments demonstrate that the method produces financially meaningful results when used for solving practical problems.

## 1 Introduction

In this paper we propose a 2nd-order numerical scheme for a 2D fractional Black-Scholes (fBS) equation arising in pricing options with two underlying assets [2], based the schemes in [4] for a 1D fBS equation. We prove that the developed discretization method is consistent, stable and monotone, and thus the solution generated by the numerical method converges to the exact one. Numerical experiments have been performed to demonstrate the order of convergence and usefulness of the scheme.

It is shown in [2] that the value of an option whose underlying asset price follows a geometric Lévy process is governed by a 1D fBS equation. Under the same assumptions, it is easy to show that the value $U$ of a two-asset option (eg. Rainbow or Basket Option) which is written on two stocks whose prices $S_1$ and $S_2$ following two independent geometric Lévy processes (with zero correlation coefficient) is determined by the following 2D fBS equation:

$$\mathcal{L}U := -U_t + a_1 U_x + a_2 U_y - b_1[{}_{-\infty}D_x^{\alpha} U] - b_2[{}_{-\infty}D_y^{\beta} U] + rU = 0 \qquad (1a)$$

for $(x, y, t) \in \mathbb{R}^2 \times [0, T)$, where $x = \ln S_1$, $y = \ln S_2$, ${}_{-\infty}D_x^{\alpha} U$ and ${}_{-\infty}D_y^{\beta} U$ denote respectively the $\alpha$-th and $\beta$-th derivatives of $U$ in $x$ and $y$ for $\alpha, \beta \in (1, 2)$, $T > 0$ is the terminal time, $r \geq 0$ is the risk-free rate, $\sigma > 0$ is the volatility of the underlying asset prices, and $a_1 = -r - \frac{1}{2}\sigma^{\alpha} \sec\left(\frac{\alpha\pi}{2}\right)$, $b_1 = a_1 + r$, $a_2 = -r - \frac{1}{2}\sigma^{\beta} \sec\left(\frac{\beta\pi}{2}\right)$, and $b_2 = a_2 + r$. In computation, the domain $\mathbb{R}^2$ has

© Springer International Publishing AG 2017
I. Dimov et al. (Eds.): NAA 2016, LNCS 10187, pp. 46–57, 2017.
DOI: 10.1007/978-3-319-57099-0_5
DISTRIBUTION A. Approved for public release: distribution unlimited.

to be truncated into $\Omega = (x_{\min}, x_{\max}) \times (y_{\min}, y_{\max})$ satisfying $x_{\min}, y_{\min} < 0$ and $x_{\max}, y_{\max} > 0$. We impose the following boundary and initial conditions

$$U(x,y,t) = U_0(x,y,t), (x,y) \in \partial\Omega,\ U(x,y,T) = U^*(x,y), \tag{1b}$$

where $U_0$ and $U^*$, satisfying $U_0(x,y,T) = U^*(x,y)$ for $(x,y) \in \partial\Omega$, are known functions depending on the types of option and the strike prices $K$ of the options. Using the aforementioned logarithmic forms, it is easy to show that $\lim_{x\to-\infty} U_x = 0$ and $\lim_{y\to-\infty} U_y = 0$ [4]. Thus, when $x_{\min}$ and $y_{\min}$ are sufficiently small, the fractional derivatives in (1a) become, up to a truncation error, the following Caputo's type

$$({}_{x_{\min}}D_x^{\alpha}, {}_{y_{\min}}D_y^{\beta})^{\top} V = \left( \int_{x_{\min}}^{x} \frac{V_{xx}(\xi,y,t)}{\Gamma_\alpha \cdot (x-\xi)^{\alpha-1}} d\xi, \int_{y_{\min}}^{y} \frac{V_{yy}(x,\xi,t)}{\Gamma_\beta \cdot (y-\xi)^{\beta-1}} d\xi \right)^{\top},$$

where $\Gamma_u = 1/\Gamma(2-u)$. In what following we will omit the subscripts $x_{\min}$ and $y_{\min}$ in the above derivative representations. Also, for any $\zeta = (\zeta_1, \zeta_2) \in (0,1]^2$, we use $\nabla^\zeta U = \left(D_x^{\zeta_1} U, D_y^{\zeta_2} U\right)^{\top}$ to denote the $\zeta$-th order gradient operator, where the fractional derivatives are of the Caputo type.

## 2 Solvability

We first reformulate (1a)–(1b) as a variational problem, and then show that the variational problem has a unique solution. Before starting this discussion, we introduce some function spaces. For any $\zeta = (\zeta_1, \zeta_2)$ and $\zeta_1, \zeta_2 \in (0,1]$, we let $H^\zeta(\Omega) := \left\{ v : v, \nabla^\zeta v \in (L^2(\Omega))^2 \right\}$. Define $|\cdot|_\zeta$ and $\|\cdot\|_\zeta$ by $|v|_\zeta^2 = \|\nabla^\zeta v\|_{L^2(\Omega)}^2$ and $\|u\|_\zeta^2 = \|u\|_{L^2(\Omega)}^2 + |u|_\zeta^2$. Then it is easy to show that $|\cdot|_\zeta$ and $\|\cdot\|_\zeta$ are seminorm and norm on $H^\zeta(\Omega)$ respectively. It has been shown in [7], that $H^\zeta(\Omega)$ equipped with $\|\cdot\|_\zeta$ is a Sobolev space. We also define the Sobolev space of functions the homogeneous boundary trace by $H_0^\zeta(\Omega) = \left\{ v : v \in H^\zeta(\Omega), v|_{\partial\Omega} = 0 \right\}$.

Without loss of generality, we assume that $U_0$ defined in (1b) satisfies $U_0 \in H^\gamma(\Omega)$, where $\gamma = (\alpha, \beta)$. Then, under the transformation $V = U_0 - U$, (1a) can be written as the following equation with boundary and payoff conditions:

$$\mathcal{L}V := -V_t - \nabla \cdot (B\nabla^{(\gamma-1)} V - aV) + rV = f, \tag{2a}$$

$$V = 0 \text{ on } \partial\Omega,\ V = V^*(x,y) := U_0(x,y,T) - U^*(x,y), \tag{2b}$$

where $a = (a_1, a_2)^{\top}$, $B = \mathrm{diag}(b_1, b_2)$, $\gamma - 1 := (\alpha - 1, \beta - 1)$, and $f(x,y,t) = \mathcal{L}U_0$. Using the notation defined above, we pose the following problem:

*Problem 1.* Find $u(t) \in H_0^{\gamma/2}(\Omega)$, such that, for all $v \in H_0^{\gamma/2}(\Omega)$,

$$\left\langle -\frac{\partial u(t)}{\partial t}, v \right\rangle + A(u(t), v) = (f(t), v)$$

almost everywhere $(a.e)$ in $(0,T)$ satisfying terminal condition (2b) a.e. in $\Omega$, where $A(u,v) = a\left\langle \nabla u, v \right\rangle + \left\langle B\nabla^{(\gamma-1)} u, \nabla v \right\rangle + r(u,v)$ with $\langle\cdot\rangle$ denoting a duality of a pair of dual spaces.

**DISTRIBUTION A. Approved for public release: distribution unlimited.**

It is easy to verify that Problem 1 is the variational problem of (2a)–(2b) (cf. [7]). From Lemma 2.1 in [4], we have shown that in the 1D case $A(\cdot,\cdot)$ is coercive and continuous. Using the lemma we now prove that $A(\cdot,\cdot)$ is also coercive and continuous, as given in the following lemma:

**Lemma 1.** *There exists a positive constant $C$, such that for any $v, w \in H_0^{\gamma/2}(\Omega)$, and $t \in (0,T)$ a.e. $A(v,v) \geq C\|v\|_{\gamma/2}^2$ and $A(v,w) \leq C\|v\|_{\gamma/2}\|w\|_{\gamma/2}$.*

The proof of this lemma, based on Lemma 2.1 in [4], is trivial and thus omitted. Using this lemma, we have the following result.

**Theorem 1.** *There exists a unique solution to Problem 1.*

This theorem is a consequence of Lemma 1 and Theorem 1.33 in [9], in which the unique solvability for an abstract variational inequality problem is established. The proof to Theorem 1 is thus omitted here.

## 3 Discretization

Numerical solution of standard BS equations has been discussed extensively in the open literature [11–14,19,21,23,24]. However, there is a very limited work available on the numerical solution of spatial fBS equations [4,10]. Various discretization schemes have been developed for fractional DEs such as those in [8,15–18]. In this section we will present a 2nd-order scheme for (1a), based on that in [4] for a 1D fBS equation.

For given positive integers $M_x$ and $M_y$, let $\Omega$ be divided into rectangular meshes with nodes $(x_i, y_j)$, $i = 0, .., M_x$, $j = 0, \ldots, M_y$, where $x_i = x_{\min} + ih_1$ and $y_j = y_{\min} + jh_2$ with $h_1 = (x_{\max} - x_{\min})/M_x$ and $h_2 = (y_{\max} - y_{\min})/M_y$. For a positive integer $N$, let $(0,T)$ be divided into $N$ sub-intervals with the mesh points $t_n = T - n\Delta t, n = 0, 1, \ldots, N$, where $\Delta t = T/N$. The $\alpha$-th partial derivative can be approximated as follows [4]:

$$D_x^\alpha V(x_i, y_j) \approx \frac{{h_1}^{-\alpha}}{\Gamma(2-\alpha)} \sum_{k=0}^{i+1} g_k^\alpha V_{i-k+1,j} \tag{3}$$

for any $i \in \{1, 2, \ldots, M_x - 1\}$ and $j \in \{1, 2, \ldots, M_y - 1\}$, where $V_{i-k+1,j}$ is an approximation to $V(x_{i-k+1}, y_j, t)$ and $g_k^\alpha$'s are given by, for $k = 3, 4, \ldots, i+1$,

$$g_0^\alpha = \frac{1}{(2-\alpha)(3-\alpha)},\ g_1^\alpha = \frac{2^{3-\alpha} - 4}{(2-\alpha)(3-\alpha)},\ g_2^\alpha = \frac{3^{3-\alpha} - 4 \times 2^{3-\alpha} + 6}{(2-\alpha)(3-\alpha)}, \tag{4}$$

$$g_k^\alpha = g_0^\alpha[(k+1)^{3-\alpha} - 4k^{3-\alpha} + 6(k-1)^{3-\alpha} - 4(k-2)^{3-\alpha} + (k-3)^{3-\alpha}]. \tag{5}$$

**Lemma 2.** *For any $\alpha \in (1,2)$, the coefficients $g_k^\alpha$, $k = 0, 1, \ldots, i+1$ satisfy:*

*(1) $g_0^\alpha > 0$, $g_1^\alpha < 0$, and $g_k^\alpha > 0$ for $k = 3, 4, 5, \ldots, i+1$,*
*(2) there exists an $\alpha^* \in (1,2)$ such that $g_2^\alpha < 0$ when $\alpha \in (1, \alpha^*)$ and $g_2^\alpha > 0$ when $\alpha \in (\alpha^*, 2)$, and*

DISTRIBUTION A. Approved for public release: distribution unlimited.

*(3)* $\sum_{k=0}^{i+1} g_k^{\alpha} < 0$.

The proof of Lemma 2 can be found in [4]. Using (3) and its counterpart for $D_y^{\beta} V(x_i, y_j)$, we define the following operators:

$$(\delta_x^{\alpha}, \delta_y^{\beta}) U_{i,j}^n = \left( \frac{h_1^{-\alpha}}{\Gamma(2-\alpha)} \sum_{k=0}^{i+1} g_k^{\alpha} U_{i-k+1,j}^n, \frac{h_2^{-\beta}}{\Gamma(2-\beta)} \sum_{k=0}^{j+1} g_k^{\beta} U_{i,j-k+1}^n \right), \tag{6a}$$

$$\delta_x U_{i,j}^n = \frac{1}{2h_1}(U_{i+1,j}^n - U_{i-1,j}^n), \quad \delta_y U_{i,j}^n = \frac{1}{2h_2}(U_{i,j+1}^n - U_{i,j-1}^n), \tag{6b}$$

where $U_{k,l}^n$ denotes an approximation to $U(x_k, y_l, t_n)$. Using (6a)–(6b), we define the following scheme for (1):

$$\begin{aligned} &\frac{U_{i,j}^{n+1} - U_{i,j}^n}{\Delta t} + \theta \left( a_1 \delta_x U_{i,j}^{n+1} - b_1 \delta_x^{\alpha} U_{i,j}^{n+1} + a_2 \delta_y U_{i,j}^{n+1} - b_2 \delta_y^{\beta} U_{i,j}^{n+1} + r U_{i,j}^{n+1} \right) \\ &+ (1-\theta) \left( a_1 \delta_x U_{i,j}^n - b_1 \delta_x^{\alpha} U_{i,j}^n + a_2 \delta_x U_{i,j}^n - b_2 \delta_y^{\beta} U_{i,j}^n + r U_{i,j}^n \right) = 0 \end{aligned} \tag{7a}$$

for $i = 1, \ldots, M_x - 1, j = 1, \ldots, M_y - 1$, and $n = 0, \ldots, N-1$ with $\theta \in [0.5, 1]$. The boundary and payoff conditions are:

$$U_{0,j}^n = U_0(x_0, y_j, t_n), \quad U_{M_x,j}^n = U_0(x_{M_x}, y_j, t_n), \quad U_{i,0}^n = U_0(x_i, y_0, t_n), \tag{7b}$$

$$U_{i,M_y}^n = U_0(x_i, y_{M_y}, t_n), \quad U_{i,j}^N = U^*(x_i, y_j, T_N) \tag{7c}$$

for all feasible $(i, j, n)$, To rewrite (7a) into a matrix form, we let

$$\mathbf{U}^n = (U_{1,1}^n, \ldots, U_{M_x-1,1}^n, U_{1,2}^n, \ldots, U_{M_x-1,1}^n, \ldots, U_{1,M_y-1,1}^n, \ldots U_{M_y-1,M_x-1}^n)^{\top}.$$

Rearranging (7a), we have

$$(\mathbf{I} + \theta \mathbf{M})\, \mathbf{U}^{n+1} = (\mathbf{I} - (1-\theta)\mathbf{M})\, \mathbf{U}^n + \mathbf{f}^{n+1-\theta}, \tag{8}$$

where $\mathbf{I}$ is $(M_x - 1)(M_y - 1)$ dimensional identity. The matrix $\mathbf{M}$ is a block matrix which has $(M_y - 1) \times (M_y - 1)$ blocks, and the size of each block matrix is $(M_x - 1) \times (M_x - 1)$.

$$\mathbf{M} = \begin{bmatrix} \mathbf{A}+\mathbf{B}_1 & \mathbf{B}_0 & \mathbf{0} & \cdots & \cdots & \cdots & \mathbf{0} \\ \mathbf{B}_2 & \mathbf{A}+\mathbf{B}_1 & \mathbf{B}_0 & \ddots & & & \mathbf{0} \\ \mathbf{B}_3 & \mathbf{B}_2 & \mathbf{A}+\mathbf{B}_1 & \mathbf{B}_0 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{B}_{M_y-3} & & \ddots & \mathbf{B}_2 & \mathbf{A}+\mathbf{B}_1 & \mathbf{B}_0 & \mathbf{0} \\ \mathbf{B}_{M_y-2} & & & \mathbf{B}_3 & \mathbf{B}_2 & \mathbf{A}+\mathbf{B}_1 & \mathbf{B}_0 \\ \mathbf{B}_{M_y-1} & \cdots & \cdots & \cdots & \mathbf{B}_3 & \mathbf{B}_2 & \mathbf{A}+\mathbf{B}_1 \end{bmatrix}_{(M_y-1)\times(M_y-1)}, \tag{9}$$

DISTRIBUTION A. Approved for public release: distribution unlimited.

where

$$A_{ij} = \begin{cases} \mu_1 g_0^\alpha + \eta_1, & j = i+1 \\ \mu_1 g_1^\alpha + \frac{r}{2}\Delta t, & j = i \\ \mu_1 g_2^\alpha - \eta_1, & j = i-1 \\ \mu_1 g_k^\alpha, & j = i-k+1 \\ & k = 3, 4, \ldots, i \\ 0, & \text{otherwise} \end{cases}, \quad \mathbf{B}_j = \begin{cases} (\mu_2 g_0^\beta + \eta_2)\mathbf{I}_y, & j = 0 \\ (\mu_2 g_1^\beta + \frac{r}{2}\Delta t)\mathbf{I}_y, & j = 1 \\ (\mu_2 g_2^\beta - \eta_2)\mathbf{I}_y, & j = 2 \\ (\mu_2 g_j^\beta)\mathbf{I}_y, & j = 3, 4, \ldots \\ & M_y - 1 \\ 0, & \text{otherwise} \end{cases},$$

$$\mu_1 = \frac{-b_1 \Delta t}{\Gamma(2-\alpha) h_1^\alpha}, \quad \eta_1 = \frac{a_1 \Delta t}{2h_1}, \quad \mu_2 = \frac{-b_2 \Delta t}{\Gamma(2-\beta) h_2^\beta}, \quad \eta_2 = \frac{a_2 \Delta t}{2h_2}, \tag{10}$$

and $\mathbf{I}_y$ is the $(M_x - 1) \times (M_x - 1)$ identity matrix. The column vector $\mathbf{f}^{n+1-\theta} = (1-\theta)\mathbf{f}^n + \theta \mathbf{f}^{n+1}$ is the contribution from the boundary conditions (7b)–(7c), where $\mathbf{f}^n$ and $\mathbf{f}^{n+1}$ consist of contributions of boundary values at $t_n$ and $t_{n+1}$ respectively. In the rest of this paper, we choose $\theta = 0.5$ which is the Crank-Nicolson method with a 2nd-order truncation error.

We comment that though the discretization method is developed for European option pricing problems, the principle developed is applicable to complementarity problems involving the fractional differential operators in (1a) governing American option valuation if a penalty method such as those in [3,14,20,22,24] is used. We will discuss this in a future paper.

## 4 Convergence Analysis

In this section, we show that the solution to (7) converges to the viscosity solution to (1). We start the discussion with the following theorem:

**Theorem 2.** *(Consistency) The finite difference scheme for (7a) is consistent with a truncation error of order* $\mathcal{O}(\Delta t^2 + h_1^2 + h_2^2)$ *when* $\theta = 0.5$.

*Proof.* In [4], we have shown that the finite difference scheme for the derivatives in $x$ in (6) have the 2nd-order truncation error $\mathcal{O}(h_1^2)$. By symmetry, the finite difference schemes in $y$-direction in (6a) and (7a) have the truncation error $\mathcal{O}(h_2^2)$. It is also known that the Crank-Nicolson's scheme used in (7) has the truncation error of order $\mathcal{O}(\Delta t^2)$. Therefore, the discretization scheme (7a) has the truncation error $\mathcal{O}(\Delta t^2 + h_1^2 + h_2^2)$.

**Theorem 3.** *(Stability) The finite difference scheme defined by (7) is unconditionally stable.*

*Proof.* we use the semi-discrete Fourier transform to prove the stability of the Crank-Nicolson method with $\theta = 1/2$. From the definition, we see that all the coefficient matrices in (9) are Toeplitz matrices. Thus, each of the terms in (9) can be written as convolution of one the following vectors with a finite

DISTRIBUTION A. Approved for public release: distribution unlimited.

support $(\cdots,0,(\mathbf{U}^k)^\top,0,\cdots)^\top$ and $(\cdots,0,(\mathbf{f}^{n+1/2})^\top,0,\cdots)^\top$ for $k=n$ and $n+1$. Applying the discrete Fourier transform via the semidiscrete Fourier transform pair $U_{i,j}^n = \frac{1}{(2\pi)^2}\int_{-\pi/h_2}^{\pi/h_2}\int_{-\pi/h_1}^{\pi/h_1} e^{\mathrm{i}(\xi_1 x_i+\xi_2 y_j)}\hat{U}^n(\xi)d\xi_1 d\xi_2$ and $\hat{U}^n(\xi_1,\xi_2) = h_2h_1\sum_{j=-\infty}^{\infty}\sum_{i=-\infty}^{\infty} U_{i,j}^n e^{-\mathrm{i}(\xi_1 x_i+\xi_2 y_j)}$ to (8), or equivalently replacing $U_{i,j}^k$ and $f_{i,j}^{n+1/2}$ with $\hat{U}^k e^{(i\xi_1 h_1+j\xi_2 h_2)\mathrm{i}}$ and $\hat{f}^{n+1/2}e^{(i\xi_1 h_1+j\xi_2 h_2)\mathrm{i}}$ with $\mathrm{i}=\sqrt{-1}$ for all admissible $i$, $j$ and $k=n,n+1$, we obtain a system in $\hat{U}^{n+1}$. Solving the transformed system for $\hat{U}^{n+1}$ we have

$$\hat{U}^{n+1} = \frac{2-\left[\bar{\eta}_1+\mu_1\sum_{k=0}^{i+1}g_k^\alpha e^{(1-k)\xi_1 h_1\mathrm{i}}+\bar{\eta}_2+\mu_2\sum_{k=0}^{j+1}g_k^\beta e^{(1-k)\xi_2 h_2\mathrm{i}}+r\Delta t\right]}{2+\left[\bar{\eta}_1+\mu_1\sum_{k=0}^{i+1}g_k^\alpha e^{(1-k)\xi_1 h_1\mathrm{i}}+\bar{\eta}_2+\mu_2\sum_{k=0}^{j+1}g_k^\beta e^{(1-k)\xi_2 h_2\mathrm{i}}+r\Delta t\right]}\hat{U}^n + \frac{2\Delta t\hat{f}^{n+1/2}}{2+\left[\bar{\eta}_1+\mu_1\sum_{k=0}^{i+1}g_k^\alpha e^{(1-k)\xi_1 h_1\mathrm{i}}+\bar{\eta}_2+\mu_2\sum_{k=0}^{j+1}g_k^\beta e^{(1-k)\xi_2 h_2\mathrm{i}}+r\Delta t\right]},$$

where $\bar{\eta}_1=\eta_1\left(e^{\xi_1 h_1\mathrm{i}}-e^{-\xi_1 h_1\mathrm{i}}\right)$, $\bar{\eta}_2=\eta_2\left(e^{\xi_2 h_2\mathrm{i}}-e^{-\xi_2 h_2\mathrm{i}}\right)$, $\xi_1\in[-\pi/h_1,\pi/h_1]$, $\xi_2\in[-\pi/h_2,\pi/h_2]$ and $\mu_1,\mu_2,\eta_1,\eta_2$ are defined in (10). Using Euler's formula, we rewrite the above equality as follows.

$$\hat{U}^{n+1} = \frac{1-[(A_1+A_2)+(B_1+B_2)\mathrm{i}]}{1+[(A_1+A_2)+(B_1+B_2)\mathrm{i}]}\hat{U}^n + \frac{\Delta t}{1+[(A_1+A_2)+(B_1+B_2)\mathrm{i}]}\hat{f}^{n+\frac{1}{2}},$$

where

$$A_1 = \frac{\mu_1}{2}\sum_{k=0}^{i+1} g_k^\alpha\cos((1-k)\xi_1 h_1)+\frac{r\Delta t}{4},$$

$$B_1 = \frac{\eta_1\sin(\xi_1 h_1)}{2}+\frac{\mu_1}{2}\sum_{k=0}^{i+1} g_k^\alpha\sin((1-k)\xi_1 h_1),$$

and $A_2$ and $B_2$ are defined by replacing the superscript-subscript pair $(\alpha,1)$ with $(\beta,2)$. Taking magnitudes on both sides of the above equation, we have

$$|\hat{U}^{n+1}| = |\hat{U}^n|\sqrt{\frac{(1-A)^2+B^2}{(1+A)^2+B^2}}+|\hat{f}^{n+\frac{1}{2}}|\frac{\Delta t}{\sqrt{(1+A)^2+B^2}}, \tag{11}$$

where $A=A_1+A_2$ and $B=B_1+B_2$. We now show that $\frac{(1-A)^2+B^2}{(1+A)^2+B^2}\le 1$, or $A>0$. Omitting the superscripts, we have from Item 3 of in Lemma 2 that $-g_1\ge\sum_{k=0,k\neq 1}^{i+1} g_k$, with $g_k>0$ when $k>3$ for all $i>3$. From the representations of $g_k$ in (4)–(5), we have that $g_0+g_2>0$. In order to estimate $A_1$ and $A_2$, we first derive the following estimate

$$\sum_{k=0}^{i+1} g_k\cos((1-k)\xi h) = g_0\cos(\xi h)+g_1\cos 0+g_2\cos(-\xi h)+\sum_{k=3}^{i+1} g_k\cos((k-1)\xi h)$$

$$= g_1+(g_0+g_2)\cos(\xi h)+\sum_{k=3}^{i+1} g_k\cos((k-1)\xi h)\le\sum_{k=0}^{i+1} g_k\le 0.$$

DISTRIBUTION A. Approved for public release: distribution unlimited.

Since $\mu_1, \mu_2 < 0$, we have the following estimations

$$\frac{\mu_1}{2}\sum_{k=0}^{i+1} g_k^{\alpha}\cos\left((1-k)\xi h_1\right) + \frac{r\Delta t}{4} \geq 0,\ \frac{\mu_2}{2}\sum_{k=0}^{i+1} g_k^{\beta}\cos\left((1-k)\xi h_2\right) + \frac{r\Delta t}{4} \geq 0.$$

Therefore, $A_1, A_2 \geq 0$ and so $A \geq 0$. Using this result we have from (11) that, for all $\xi_i \in [-\frac{\pi}{h_i}, \frac{\pi}{h_i}]$, $i = 1, 2$,

$$\begin{aligned}\left|\hat{U}^{n+1}\right| &\leq \left|\hat{U}^{n}\right| + \Delta t\left|\hat{f}^{n+1/2}\right| \leq \left|\hat{U}^{n-1}\right| + \Delta t\left[\left|\hat{f}^{n+1/2}\right| + \left|\hat{f}^{n-1/2}\right|\right] \\ &\leq \cdots \leq \left|\hat{U}^{0}\right| + \Delta t\sum_{k=0}^{n}\left|\hat{f}^{k+1/2}\right| \leq \left|\hat{U}^{0}\right| + \frac{T}{N}\sum_{k=0}^{n}\left|\hat{f}^{k+1/2}\right|.\end{aligned}$$

Using Cauchy-Schwarz inequality, we have

$$\left|\hat{U}^{n+1}\right|^2 \leq C\left(\left|\hat{U}^{0}\right|^2 + \frac{nT^2}{N^2}\sum_{k=0}^{n}\left|\hat{f}^{k+1/2}\right|^2\right) \leq C\left(\left|\hat{U}^{0}\right|^2 + \frac{1}{N}\sum_{k=0}^{n}\left|\hat{f}^{k+1/2}\right|^2\right)$$

for any $n \leq N-1$, where $C$ denotes a generic positive constant, independent of $n$ and $N$, $\hat{U}^{n+1}, \hat{U}^0$ and $\hat{f}^{k+1/2}$ are all functions of $\xi_i \in [-\frac{\pi}{h_i}, \frac{\pi}{h_i}]$ for $i = 1, 2$. For any continuous function $W$ on $\bar{\Omega}$, let $||W||_{0,h} = \left(h_2 h_1 \sum_{j=1}^{M_y-1}\sum_{i=1}^{M_x-1}|W_{i,j}|^2\right)^{1/2}$ denote the discrete $L^2$-norm of $W$. Using the properties of the discrete Fourier and its inverse transforms (particularly Parseval's equality) we have

$$\begin{aligned}&||\mathbf{U}^{j+1}||_{0,h}^2 = \frac{1}{(2\pi)^2}\int_{-\pi/h_2}^{\pi/h_2}\int_{-\pi/h_1}^{\pi/h_1}|\hat{U}^{n+1}|^2 d\xi_1 d\xi_2 \\ &\leq \frac{1}{(2\pi)^2}\left(\int_{-\pi/h_2}^{\pi/h_2}\int_{-\pi/h_1}^{\pi/h_1}|\hat{U}^{0}|^2 d\xi_1 d\xi_2 + \frac{T}{N}\sum_{k=0}^{n}\int_{-\pi/h_2}^{\pi/h_2}\int_{-\pi/h_1}^{\pi/h_1}|\hat{f}^{k+1/2}|^2 d\xi_1 d\xi_2\right) \\ &= C\left(||\mathbf{U}^0||_{0,h}^2 + \frac{1}{N}\sum_{k=0}^{n}||\mathbf{f}^{k+1/2}||_{0,h}^2\right) \leq C\left(||\mathbf{U}^0||_{0,h}^2 + ||\mathbf{f}||_{\infty}^2\right),\end{aligned}$$

where $\mathbf{f} = ((\mathbf{f}^0)^\top, \ldots, (\mathbf{f}^N)^\top)^\top$. Thus, we obtain $||\mathbf{U}^{j+1}||_{0,h} \leq C\left(||\mathbf{U}^0||_{0,h} + ||\mathbf{f}||_{\infty}\right)$. Therefore, the numerical method is unconditionally stable.

We now show that the numerical scheme is monotone.

**Theorem 4.** *(Monotonicity) The discretization scheme established in (7) is monotone when $\Delta t \leq \frac{2}{r}$.*

*Proof.* By rearranging the discretizated equation (7a), we define a linear function $F_{i,j}^n$ of $U^{n+1}$ and $U^n$ as follows:

DISTRIBUTION A. Approved for public release: distribution unlimited.

$$
\begin{aligned}
&2F_{i,j}^{n+1}\big(U_{i,j}^{n+1}, U_{i-1,j}^{n+1}, \ldots, U_{0,j}^{n+1}, U_{i,j+1}^{n+1}, U_{i,j-1}^{n+1}, \ldots, U_{i,0}^{n+1}, U_{i+1,j}^{n}, U_{i,j}^{n},\\
&U_{i-1,j}^{n}, \ldots, U_{0,j}^{n}, U_{i,j+1}^{n}, U_{i,j-1}^{n}, \ldots, U_{i,0}^{n}\big) := \left[2 + \left(\mu_1 g_1^{\alpha} + \mu_2 g_1^{\beta} + r\Delta t\right)\right] U_{i,j}^{n+1}\\
&+ \left(\eta_1 + \mu_1 g_0^{\alpha}\right) U_{i+1,j}^{n+1} - \left(\eta_1 - \mu_1 g_2^{\alpha}\right) U_{i-1,j}^{n+1} + \mu_1 \sum_{k=3}^{i+1} g_k^{\alpha} U_{i-k+1,j}^{n+1} + \left(\eta_2 + \mu_2 g_0^{\beta}\right) U_{i,j+1}^{n+1}\\
&- \left(\eta_2 - \mu_2 g_2^{\beta}\right) U_{i,j-1}^{n+1} + \mu_2 \sum_{k=3}^{j+1} g_k^{\beta} U_{i,j-k+1}^{n+1} - \left[2 - \left(\mu_1 g_1^{\alpha} + \mu_2 g_1^{\beta} + r\Delta t\right)\right] U_{i,j}^{n}\\
&+ \left(\eta_1 + \mu_1 g_0^{\alpha}\right) U_{i+1,j}^{n} - \left(\eta_1 - \mu_1 g_2^{\alpha}\right) U_{i-1,j}^{n} + \mu_1 \sum_{k=3}^{i+1} g_k^{\alpha} U_{i-k+1,j}^{n}\\
&+ \left(\eta_2 + \mu_2 g_0^{\beta}\right) U_{i,j+1}^{n} - \left(\eta_2 - \mu_2 g_2^{\beta}\right) U_{i,j-1}^{n} + \mu_2 \sum_{k=3}^{j+1} g_k^{\beta} U_{i,j-k+1}^{n}.
\end{aligned}
$$

We also define the following two functions:

$$
\begin{aligned}
F_{i,j,+\varepsilon}^{n+1} &= F_{i,j}^{n+1}\big(U_{i,j}^{n+1} + \varepsilon, U_{i+1,j}^{n+1}, U_{i-1,j}^{n+1}, \ldots, U_{0,j}^{n+1}, \ldots, U_{i,j+1}^{n}, U_{i,j-1}^{n}, \ldots, U_{i,0}^{n}\big)\\
F_{i,j,-\varepsilon}^{n+1} &:= F_{i,j}^{n+1}\big(U_{i,j}^{n+1}, U_{i+1,j}^{n+1} + \varepsilon, U_{i-1,j}^{n+1} + \varepsilon, \ldots, U_{0,j}^{n+1} + \varepsilon, \ldots, U_{i,j+1}^{n} + \varepsilon,\\
&\qquad U_{i,j-1}^{n} + \varepsilon, \ldots, U_{i,0}^{n} + \varepsilon\big),
\end{aligned}
$$

where $\varepsilon > 0$. It has been proved in [4] that $\left(\sum_{k=0}^{i+1} g_k^{\alpha}\right) - \frac{1}{2} g_1^{\alpha} > 0$ for $i = 1, 2, \ldots, M_x - 1$. This inequality also holds true for $\{g_k^{\beta}\}$ with $i+1$ and $M_x$ replaced with $j+1$ and $M_y$ respectively. We now use this result to prove the monotonicity of $F_{i,j}^{n+1}$. When $\Delta t \le \frac{2}{r}$, we have from the definition of $F_{i,j}^{n+1}$ that, for any $\varepsilon > 0$ and feasible $i$ and $j$,

$$
\begin{aligned}
F_{i,j,-\varepsilon}^{n+1} &= F_{i,j}^{n+1} - \left[1 - \frac{1}{2}\left(\mu_1 g_1^{\alpha} + \mu_2 g_1^{\beta} + r\Delta t\right)\right]\varepsilon + \mu_1 (g_0^{\alpha} + g_2^{\alpha})\varepsilon\\
&+ \mu_1 \sum_{k=3}^{i+1} g_k^{\alpha}\varepsilon + \mu_2 (g_0^{\beta} + g_2^{\beta})\varepsilon + \mu_2 \sum_{k=3}^{j+1} g_k^{\beta}\varepsilon\\
&\le F_{i,j}^{n+1} + \mu_1 \left(\sum_{k=0}^{i+1} g_k^{\alpha} - \frac{1}{2} g_1^{\alpha}\right)\varepsilon + \mu_2 \left(\sum_{k=0}^{j+1} g_k^{\beta} - \frac{1}{2} g_1^{\beta}\right)\varepsilon - \left(1 - \frac{1}{2} r\Delta t\right)\varepsilon \le F_{i,j}^{n+1},
\end{aligned}
$$

since $\mu_1, \mu_2 < 0$. Furthermore, from Lemma 2, we know that $g_1^{\alpha} < 0$ and $g_1^{\beta} < 0$, thus we have

$$
F_{i,j,+\varepsilon}^{n+1} = F_{i,j}^{n+1} + \left[1 + \frac{1}{2}\left(\mu_1 g_1^{\alpha} + \mu_2 g_1^{\beta} + r\Delta t\right)\right]\varepsilon > F_{i,j}^{n+1}.
$$

Therefore, the scheme is monotone.

Combining Theorems 2, 3 and 4, we have the following convergence result.

**DISTRIBUTION A. Approved for public release: distribution unlimited.**

**Theorem 5.** *(Convergence) Let $U$ be the viscosity solution to (1) and $U_{h_1,h_2,\Delta t}$ be the numerical solution to (7) with spatial and time mesh size triple $(h_1, h_2, \Delta t)$. Then, $U_{h_1,h_2,\Delta t}$ converges to $U$ as $(h_1, h_2, \Delta t) \to (0^+, 0^+, 0^+)$.*

In [1] the authors show that any finite difference scheme for a general nonlinear 2nd-order PDE which is locally consistent, stable and monotone generates a solution converging uniformly on a compact subset of $(0, T) \times \mathbb{R}$ to the unique viscosity solution of the PDE. In [5,6], Cont and Tankov extended this result to partial integro-differential equations (PIDEs). Since (1a) is an PIDE, Theorem 5 is a consequence of the results established in [1,5,6] and Theorems 2, 3 and 4.

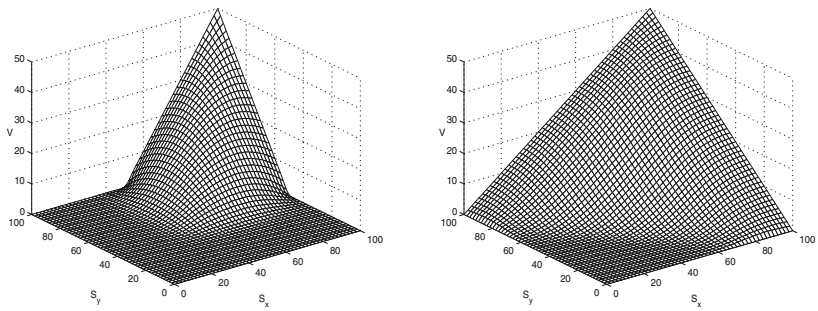
## 5 Numerical Experiments

We now apply our method to the following test problem.

*Example 1.* Call-on-Min and Basket options: Eq. (1), with the system and market parameters $\sigma = 0.25$, $r = 0.05$, $K = 50$, $a_1 = a_2 = 0.384$, $b_1 = b_2 = 0.884$, $x_{\min} = y_{\min} = \ln(0.1)$, $x_{\max} = y_{\max} = \ln(100)$ and $T = 1$. Initial and boundary conditions can be obtained by setting $t = T$, $x = x_{\min}, x_{\max}$ or $y = y_{\min}, y_{\max}$ in the following functions.

Call-on-Min option: $U(x, y, t) = \left[\min(e^x, e^y) - Ke^{-r(T-t)}\right]_+$;
Basket option: $U(x, y, t) = \left[(e^x + e^y)/2 - Ke^{-r(T-t)}\right]_+$.

To solve this problem, we choose a uniform mesh with mesh sizes $\Delta x = \Delta y = \frac{1}{100}$ and $\Delta t = \frac{1}{100}$. The numerical solutions for these options at $t = 0$ from our method are plotted in Fig. 1 in the original independent variable $S_x = e^x$ and $S_y = e^y$. From the figures we see that these numerical solutions are qualitatively correct.

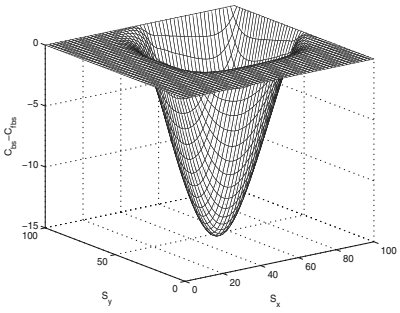



Call-On-Min Basket

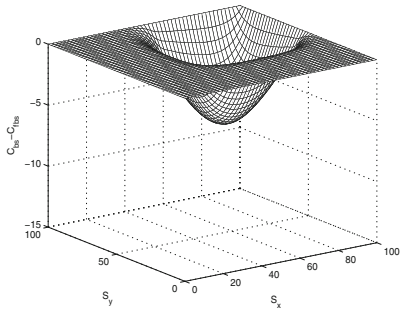
**Fig. 1.** Computed prices of Call-on-Min and Basket options; $\alpha = \beta = 1.5$

DISTRIBUTION A. Approved for public release: distribution unlimited.

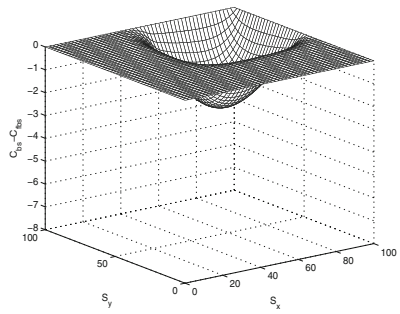
To see the influence of $\alpha$ and $\beta$ on the option prices, we solve the problem for four different values of $\alpha = \beta = 1.3, 1.5, 1.7, 1.9$ and plot the differences between the numerical solutions of the standard BS equation (i.e., $\alpha = \beta = 2$) and the fractional BS equation and at $t = 0$ for Call-on-Min (Fig. 2) and Basket Option (Fig. 3). From the figures we see that the Call-on-Min and Basket options from fBS model are more expensive than their counterparts of the standard BS model. From these figures, we also see that the call prices increase as $\alpha$ decreases when $S_1$ and $S_2$ are greater than some critical values. This phenomenon has been observed in published results for of the 1D fBS equation [2,4] and thus our numerical results for the 2D problem are consistent with those from [2]. The figures also indicate that when $\alpha$ and $\beta$ approach 2, the numerical solutions to the fBS equation approach to those of the BS equation.



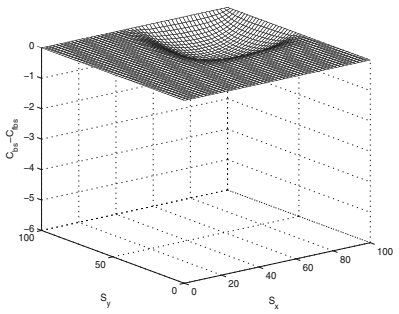

$\alpha, \beta = 1.3$




$\alpha, \beta = 1.5$




$\alpha, \beta = 1.7$




$\alpha, \beta = 1.9$

**Fig. 2.** $V_{bs} - V_{fbs}$ Call-on-Min option

**DISTRIBUTION A. Approved for public release: distribution unlimited.**

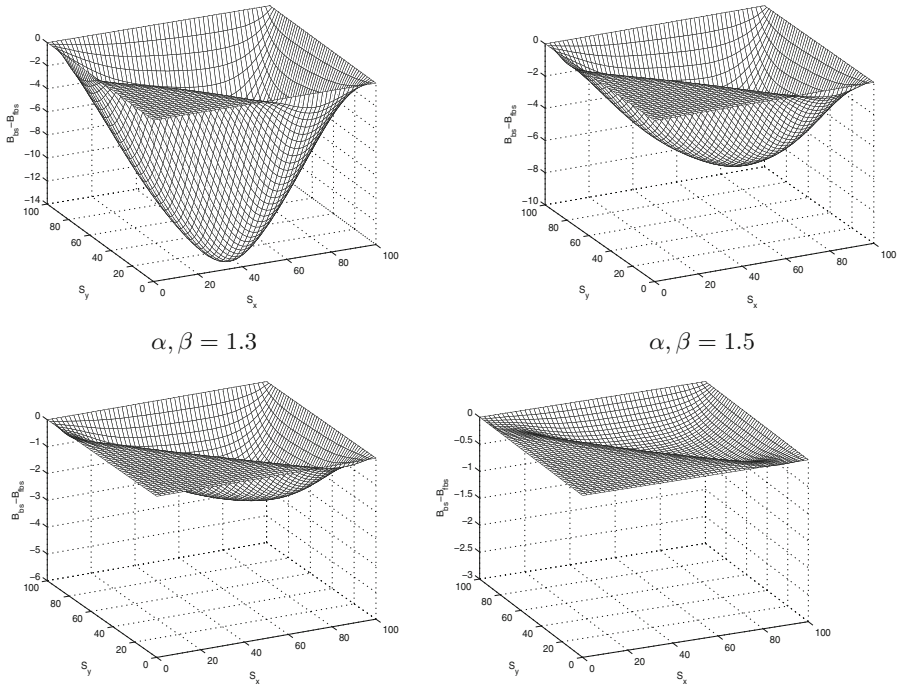



**Fig. 3.** $V_{bs} - V_{fbs}$ Basket option

## 6 Conclusion

In this paper, an FDM is proposed to solve the 2D fractional Black-Scholes equation. The discretization method is shown to be unconditionally stable and convergent. Numerical experiments are performed to demonstrate the usefulness of the methods for pricing two-asset European options of practical significance.

**Acknowledgements.** S. Wang's work was partially supported by the AOARD Project #15IOA095.

## References

1. Barles, G.: Convergence of numerical schemes for degenerate parabolic equations arising in finance theory. In: Rogers, L.C.G., Talay, D. (eds.) Numerical Methods in Finance. Cambridge University Press, Cambridge (1997)
2. Cartea, A., del-Castillo-Negrete, D.: Fractional diffusion models of option prices in markets with jumps. Phys. A **374**, 749–763 (2007)
3. Chen, W., Wang, S.: A penalty method for a fractional order parabolic variational inequality governing American put option valuation. Comput. Math. Appl. **67**, 77–90 (2014)
4. Chen, W., Wang, S.: A finite difference method for pricing European and American options under a geometric Lévy process. J. Ind. Manag. Optim. **11**, 241–264 (2015)
5. Cont, R., Tankov, P.: Financial Modelling with Jump Processes, vol. 2. Chapman & Hall, New York (2004)

**DISTRIBUTION A. Approved for public release: distribution unlimited.**

6. Cont, R., Voltchkova, E.: A finite difference scheme for option pricing in jump-diffusion and exponential Lévy models. Ecole Polytechnique Rapport Interne CMAP Working Paper No. 513 (2005)
7. Ervin, V.J., Roop, J.P.: Variational formulation for the stationary fractional advection dispersion equation. Numer. Methods Partial Differ. Equ. **22**, 558–576 (2006)
8. Ervin, V.J., Heuer, N., Roop, J.P.: Numerical approximation of a time dependent, nonlinear, space-fractional diffusion equation. SIAM J. Numer. Anal. **45**, 572–591 (2007)
9. Haslinger, J., Miettinen, M.: Finite Element Method for Hemivariational Inequalities. Kluwer Academic Publisher, Dordrecht-Boston-London (1999)
10. Koleva, M.N., Vulkov, L.G.: Numerical solution of time-fractional BlackScholes equation. Comput. Appl. Math. (2016). doi:10.1007/s40314-016-0330-z
11. Lesmana, D.C., Wang, S.: An upwind finite difference method for a nonlinear Black-Scholes equation governing European option valuation. Appl. Math. Comput. **219**, 8818–8828 (2013)
12. Lesmana, D.C., Wang, S.: Penalty approach to a nonlinear obstacle problem governing American put option valuation under transaction costs. Appl. Math. Comput. **251**, 318–330 (2015)
13. Li, W., Wang, S.: Penalty approach to the HJB equation arising in European stock option pricing with proportional transaction costs. J. Optim. Theory Appl. **143**, 279–293 (2009)
14. Li, W., Wang, S.: Pricing American options under proportional transaction costs using a penalty approach and a finite difference scheme. J. Ind. Manag. Optim. **9**, 365–389 (2013)
15. Lynch, V.E., Carreras, B.A., del-Castillo-Negrete, D., Ferreira-Mejias, K.M., Hicks, H.R.: Numerical methods for the solution of partial differential equations of fractional order. J. Comput. Phys. **192**, 406–421 (2003)
16. Meerschaert, M.M., Tadjeran, C.: Finite difference methods for two-dimensional fractional dispersion equation. J. Comput. Phys. **211**, 249–261 (2006)
17. Oldham, K.B., Spanier, J.: The Fractional Calculus. Academic Press, Cambridge (1974)
18. Tadjeran, C., Meerschaert, M.M.: A second-order accurate numerical method for the two-dimensional fractional diffusion equation. J. Comput. Phys. **220**, 813–823 (2007)
19. Wang, S.: A novel fitted finite volume method for the Black-Scholes equation governing option pricing. IMA J. Numer. Anal. **24**, 699–720 (2004)
20. Wang, S., Yang, X.Q., Teo, K.L.: Power penalty method for a linear complementarity problem arising from American option valuation. J. Optim. Theory Appl. **129**(2), 227–254 (2006)
21. Wang, S., Zhang, S., Fang, Z.: A superconvergent fitted finite volume method for BlackScholes governing European and American option valuation. Numer. Methods Partial Differ. Equ. **31**, 1190–1208 (2015)
22. Wang, S., Zhang, K.: An interior penalty method for a finite-dimensional linear complementarity problem in financial engineering. Optim. Lett. (2016). doi:10.1007/s11590-016-1050-4
23. Wilmott, P., Dewynne, J., Howison, S.: Option Pricing: Mathematical Models and Computation. Oxford Financial Press, Oxford (1993)
24. Zhang, K., Wang, S.: Pricing American bond options using a penalty method. Automatica **48**, 472–479 (2012)

**DISTRIBUTION A. Approved for public release: distribution unlimited.**

Applied Mathematics and Computation 305 (2017) 174–187

Contents lists available at ScienceDirect

Applied Mathematics and Computation

journal homepage: www.elsevier.com/locate/amc

# A power penalty method for a 2D fractional partial differential linear complementarity problem governing two-asset American option pricing

Wen Chen[a], Song Wang[b,*]

[a] CSIRO Data61, 34 Village Street, Docklands, Vic, Australia
[b] Department of Mathematics & Statistics, Curtin University, GPO Box U1987, Perth, Australia

ARTICLE INFO

*Keywords:*
American option pricing
Optimal control
Linear complementarity problem
Fractional differential equation
Penalty method
Finite difference method

ABSTRACT

In this paper we propose a power penalty method for a linear complementarity problem (LCP) involving a fractional partial differential operator in two spatial dimensions arising in pricing American options on two underlying assets whose prices follow two independent geometric Lévy processes. We first approximate the LCP by a nonlinear 2D fractional partial differential equation (fPDE) with a penalty term. We then prove that the solution to the fPDE converges to that of the LCP in a Sobolev norm at an exponential rate depending on the parameters used in the penalty term. The 2D fPDE is discretized by a 2nd-order finite difference method in space and Crank–Nicolson method in time. Numerical experiments on a model Basket Option pricing problem were performed to demonstrate the convergent rates and the effectiveness of the penalty method.

© 2017 Elsevier Inc. All rights reserved.

## 1. Introduction

Option valuation through a partial differential equation approach has been increasingly attracting much attention from financial engineers, mathematicians and statisticians, ever since the publication of the two seminal papers [4] and [20]. In [4] the authors showed that in a complete market the price of an option on a stock whose price follows a geometric Brownian motion with constant drift and volatility satisfies a second order parabolic partial differential equation, known as the Black–Scholes (BS) equation. However, Gaussian shocks used in BS model often underestimate the probability that stock prices usually exhibit large movements over small time steps which can be demonstrated by empirical financial market data. When jumps are large and rare, a jump-diffusion pricing model can be used to capture them. More details of these models and their numerical solutions can be found in, for example, [1,2,13,30,31]. If there are infinitely many jumps in a finite time interval, an infinite activity Lévy process can be used to capture both frequent small and rare large moves. It has been shown in [6] that, the price of an option on a single asset satisfies a 1D parabolic fractional Black–Scholes (fBS) equation when its underlying asset price follows a geometric Lévy process. This 1D fBS equation and the corresponding American option pricing problem can be solved numerically by the numerical methods proposed recently by us in [7,8]. In [10], Clift and Forsyth proposed an implicit finite difference method for the two dimensional parabolic partial

* Corresponding author.
*E-mail addresses:* Wen.Chen@data61.csiro.au (W. Chen), song.wang@curtin.edu.au, songwang58@gmail.com (S. Wang).

http://dx.doi.org/10.1016/j.amc.2017.01.069
0096-3003/© 2017 Elsevier Inc. All rights reserved.

DISTRIBUTION A. Approved for public release: distribution unlimited.

integro-differential equation (PIDE) to price two-asset European and American options whose assets follow the correlated finite activity jump diffusion model.

In this work, we shall present a numerical method consisting of a penalty approach and a discretization scheme for pricing American options written on two assets whose prices follow two independent geometric Lévy processes. Under the same assumptions as in [6], it is easy to show that the value of such a two-asset option of European type (eg. Rainbow or Basket Option) is determined by a 2D fBS equation and the value of the corresponding American option is governed by a linear complementarity problem involving the fractional partial differential operator used in the European option model. The latter can also be formulated as a fractional partial differential variational inequality. We comment that, the CGMY jump-diffusion process [5] is also popular in option pricing. An fBS equation for pricing European options has also been developed in [6]. However, in the present work, we only consider the fBS equations and inequalities associated with the geometric Lévy process and will develop algorithms for the fractional differential LCPs based on the CGMY jump-diffusion process in a future paper.

Penalty approaches have been used very successfully for solving constrained optimization problems. In recent years, penalty methods have been used for complementarity or variational inequality problems in both finite and infinite dimensions [3,25,35], particularly those from the valuation of financial options [15,18,19,22,27,33,34,36]. Recently, modern optimization techniques such as the use of grossone theory proposed in [24] in nonlinear programming problems with differentiable penalty functions to determine the penalty parameters has been developed in [11]. In [8], we proposed a power penalty method for solving the fBS equation governing single-asset American option pricing. In this paper, we construct and analyze a power penalty method for the fractional differential complementarity problem arising in pricing the aforementioned two-asset American options. In particular, we will establish a convergence theory for the penalty method proposed. We will then propose a 2nd-order accurate scheme for the discretization of the 2D nonlinear fBS equation in two spatial dimensions generated by the penalty method, based on our recent work in [7] for the 1D fBS equation arising in pricing one-asset options.

While the numerical solution of fractional differential LCPs and fBS equations arising in pricing options written on one risky asset has been discussed in various existing works, to our best knowledge, there are no numerical methods for their 2D counterparts governing the valuation of options on two assets. Therefore, the present work will fill this gap and provide a numerical tool for pricing European and American options of the aforementioned type.

The organization of this paper is as follows. In the next section, we will give a brief account of the fBS equation and fractional differential LCP, along with their initial and boundary conditions, governing the valuation of European and American options written on two independent risky assets. We will also formulate the LCP as a variational inequality and show that the latter problem is uniquely solvable. In Section 3, we will first propose the power penalty method with positive penalty parameters $\lambda > 1$ and $k$, and consider the unique solvability of the penalty equation. We will then prove that the solution to the penalty equation converges to that of the variational inequality at the rate $\mathcal{O}(\lambda^{-k/2})$. A 2nd-order accurate discretization scheme is proposed in Section 4 for the penalty equation. In Section 5, we will present some numerical experimental results using an American Basket option pricing problem to numerically demonstrate the rates of convergence and usefulness of the numerical method.

## 2. The option pricing problem

It is shown in [6] that the value of an option whose price follows a geometric Lévy process is governed by a 1D fBS equation. Under the same assumptions as in [6], it is trivial to show that the value $U$ of a European option written on two assets (eg. Rainbow or Basket Option) whose prices $S_1$ and $S_2$ follow two independent geometric Lévy processes is determined by the following two-dimensional fBS equation:

$$\mathcal{L}U := -U_t + a_1 U_x + a_2 U_y - b_1[{}_{-\infty}D_x^{\alpha} U] - b_2[{}_{-\infty}D_y^{\beta} U] + rU = 0 \tag{1a}$$

for $(x, y, t) \in (-\infty, \infty)^2 \times [0, T)$, where $x = \ln S_1$, $y = \ln S_2$, ${}_{-\infty}D_x^{\alpha} U$ and ${}_{-\infty}D_y^{\beta} U$ denote respectively the $\alpha$th and $\beta$th derivatives of $U$ in $x$ and $y$ for $\alpha, \beta \in (1, 2)$, $T > 0$ is the expiry date, $r \geq 0$ is the risk-free rate, $\sigma > 0$ is the volatility of the underlying asset price, and

$$a_1 = -r - \frac{1}{2}\sigma^{\alpha} \sec\left(\frac{\alpha\pi}{2}\right), \quad b_1 = -\frac{1}{2}\sigma^{\alpha} \sec\left(\frac{\alpha\pi}{2}\right) > 0,$$

$$a_2 = -r - \frac{1}{2}\sigma^{\beta} \sec\left(\frac{\beta\pi}{2}\right), \quad b_2 = -\frac{1}{2}\sigma^{\beta} \sec\left(\frac{\beta\pi}{2}\right) > 0.$$

Boundary and terminal conditions can be defined for the above equation depending on the types of options and the strike price $K$. From the transformations $x = \ln S_1$ and $y = \ln S_2$, we have

$$\lim_{x \to -\infty} U_x = \lim_{x \to -\infty} U_{S_1} e^x = 0, \quad \lim_{y \to -\infty} U_y = \lim_{y \to -\infty} U_{S_2} e^y = 0,$$

since $U_{S_1}$ and $U_{S_2}$ are bounded as $S_1, S_2 \to 0^+$ in practice. In computation, the infinite solution domain $(-\infty, \infty)^2$ has to be truncated into $\Omega = (x_{\min}, x_{\max}) \times (y_{\min}, y_{\max})$, where $x_{\min}$, $x_{\max}$, $y_{\min}$ and $y_{\max}$ are four constants satisfying

DISTRIBUTION A. Approved for public release: distribution unlimited.

$x_{\min}, y_{\min} \ll 0$ and $x_{\max}, y_{\max} > 0$. Therefore, since both $U_x$ and $U_y$ go to zero exponentially as $x$ and $y$ approach $-\infty$, when $x_{\min}, y_{\min} \ll 0$, the following conditions for (1a), up to a truncation error, are satisfied:

$$U_x(x, y, t) = 0, \quad x \le x_{\min}, (y, t) \in (y_{\min}, y_{\max}) \times [0, T), \tag{1b}$$

$$U_y(x, y, t) = 0, \quad y \le y_{\min}, (x, t) \in (x_{\min}, x_{\max}) \times [0, T). \tag{1c}$$

Clearly, for put options, the strike price $K$ should satisfy $\max(e^{x_{\min}}, e^{y_{\min}}) < K < \min(e^{x_{\max}}, e^{y_{\max}})$.

As in the conventional case, the two-asset American option price satisfies the following fractional differential linear complementarity problem

$$\mathcal{L}U \ge 0, \quad U \ge U^*, \tag{2a}$$

$$\mathcal{L}U \cdot (U - U^*) = 0, \tag{2b}$$

where $U^*$ is a given function of $(x, y, t)$ defining a 'lower bound' on the solution which is usually the payoff function of the pricing problem. Note that (2a) and (2b) contain (1a) as the special case when $U^* \le 0$. This is because $U \ge 0$ and thus the 2nd inequality in (2a) is always satisfied if $U^* \le 0$. In this case, the complementarity condition (2b) yields (1a). In what follows, we only consider pricing American puts as the price of an American call is equal to that of its European counterpart. For brevity, we assume in the rest of this work that $U^*$ is the payoff function of the problem.

On the boundary of $\Omega$, we impose the following boundary and terminal conditions for an American put:

$$U(x_{\min}, y, t) = g_1(y, t), \quad U(x, y_{\min}, t) = g_2(x, t), \quad U(x_{\max}, y, t) = 0 = U(x, y_{\max}, t) \tag{2c}$$

for $(x, y, t) \in \Omega \times [0, T)$, and

$$U(x, y, T) = U^*(x, y), \quad (x, y) \in \Omega, \tag{2d}$$

where $g_1$, $g_2$ are given functions and $U^*$ is the payoff of the option defined below.

For brevity, we only consider two-asset American Basket options in this work whose pay-off function is

$$U^*(x, y) = [K - w_1 e^x - w_2 e^y]_+, \tag{3}$$

where $[z]_+ = \max\{z, 0\}$ and $w_1, w_2 \ge 0$ are weights. Clearly, the weights are arbitrary as long as (2c) and (2d) are consistent. We also assume that the computational domain $\Omega$ is sufficiently large so that $K - w_{1e}^x - w_{2e}^x = 0$ is a curve in the interior of $\Omega$.

It is normally not possible to derive explicit analytical expressions for the boundary conditions $g_1$ and $g_2$ in this case, as they are usually the solutions of one-dimensional American option pricing problems or LCPs of the form (2a) and (2b). In practice, numerical approximations to these 1D American option pricing problem are sought as discussed in [16,17].

To determine $g_1$, one needs to solve a 1D LCP obtained by taking the limit of (2) as $x \to x_{\min}$. Using (1b), we see that $g_1$ should satisfy the following 1D LCP:

$$\begin{cases} -g_{1t} + a_2 g_{1y} - b_2[{}_{-\infty}D_y^{\beta} g_1] + r g_1 \ge 0, \\ g_1 \ge U^*(x_{\min}, y), \\ \left(-g_{1t} + a_2 g_{1y} - b_2\left[{}_{-\infty}D_y^{\beta} g_1\right] + r g_1\right) \cdot (g_1 - U^*(x_{\min}, y)) = 0, \end{cases} \tag{4}$$

with the boundary and terminal conditions

$$g_1(y_{\min}, t) = U^*(x_{\min}, y_{\min}), \quad g_1(y_{\max}, t) = 0, \quad g_1(y, T) = U^*(x_{\min}, y). \tag{5}$$

According to [29], the upper bound of the asset prices are usually three to four times the strike price. Choosing a reasonable large upper bound, we can have the above artificial boundary conditions at $(y_{\max}, t)$.

Similarly, $g_2(x, t)$ is determined by the following LCP:

$$\begin{cases} -g_{2t} + a_2 g_{2y} - b_1[{}_{-\infty}D_x^{\alpha} g_2] + r g_2 \ge 0, \\ g_2 \ge U^*(x, y_{\min}), \\ \left(-g_{2t} + a_2 g_{2y} - b_1[{}_{-\infty}D_x^{\alpha} g_2] + r g_2\right) \cdot (g_2 - U^*(x, y_{\min})) = 0, \end{cases} \tag{6}$$

with the boundary and terminal conditions:

$$g_2(x_{\min}, t) = U^*(x_{\min}, y_{\min}), \quad g_2(x_{\max}, t) = 0, \quad g_2(x, T) = U^*(x, y_{\min}). \tag{7}$$

Both (4) and (6) are single-asset American option pricing problems with fractional Black–Scholes operators. Note that the boundary and payoff conditions (5) and (7) are exact. The above 1D problems can be solved numerically using the discretization and penalty methods proposed in [7,8] to yield approximations to $g_1$ and $g_2$. The computational errors in the numerical solutions of the boundary conditions are of the order $\mathcal{O}(h^2 + \Delta t^2 + \lambda^{k/2})$ as proved in [7,8], where $h$ and $\Delta t$ are respectively the maximal mesh sizes in space and time, and $\lambda > 1$ and $k > 0$ are the penalty parameter and power used in the power penalty method.

DISTRIBUTION A. Approved for public release: distribution unlimited.

There are various representations of the fractional derivative ${}_{-\infty}D_x^\alpha U(x,y)$ such as those of Riemann–Liouville and Grüwald–Letnikov [21,23]. For a given $x_0$, one form for ${}_{x_0}D_x^\alpha V(x,y)$ is

$$ {}_{x_0}D_x^\alpha U(x,y,t) = \frac{U(x_0,y,t)}{\Gamma(1-\alpha)(x-x_0)^\alpha} + \frac{U_x(x_0,y,t)}{\Gamma(2-\alpha)(x-x_0)^{\alpha-1}} + \frac{1}{\Gamma(2-\alpha)}\int_{x_0}^{x} \frac{U_{xx}(\xi,y,t)}{(x-\xi)^{\alpha-1}}d\xi \tag{8} $$

for $x > x_0$, where $\Gamma(\cdot)$ denotes the Gamma function. Using (1b) it is easily seen that, for $x_0 \le x_{\min}$, (8) reduces to

$$ {}_{x_0}D_x^\alpha U(x,y,t) = \frac{U(x_{\min},y,t)}{\Gamma(1-\alpha)(x-x_0)^\alpha} + \frac{1}{\Gamma(2-\alpha)}\int_{x_{\min}}^{x} \frac{U_{xx}(\xi,y,t)}{(x-\xi)^{\alpha-1}}d\xi, $$

since $U_x(x,y,t) = 0$ and $U(x_0,y,t) = U(x_{\min},y,t)$ when $x \le x_{\min}$ (up to a truncation error). Therefore, we have

$$ \begin{aligned} {}_{-\infty}D_x^\alpha U(x,y,t) &= \lim_{x_0\to-\infty}\left[\frac{U(x_{\min},y,t)}{\Gamma(1-\alpha)(x-x_0)^\alpha} + \frac{1}{\Gamma(2-\alpha)}\int_{x_{\min}}^{x} \frac{U_{xx}(\xi,y,t)}{(x-\xi)^{\alpha-1}}d\xi\right] \\ &= \frac{1}{\Gamma(2-\alpha)}\int_{x_{\min}}^{x} \frac{U_{xx}(\xi,y,t)}{(x-\xi)^{\alpha-1}}d\xi \end{aligned} \tag{9} $$

for $x > x_{\min}$. This is Caputo's representation of the $\alpha$th derivative of our solution $U$ with respect to $x$. Similarly, using (1c), we can derive, for $y > y_{\min}$ and up to a truncation error,

$$ {}_{-\infty}D_y^\beta U(x,y,t) = \frac{1}{\Gamma(2-\beta)}\int_{y_{\min}}^{y} \frac{U_{yy}(x,\xi,t)}{(y-\xi)^{\beta-1}}d\xi. $$

### 2.1. The variational formulation and unique solvability

In this section, we first formulate (2) as a variational inequality problem and then show that the problem has a unique solution. We start this discussion by introducing some function spaces.

For the open set $\Omega \subseteq \mathbb{R}^2$ and $1 \le p \le \infty$, we let $L^p(\Omega) = \{v : (\int_\Omega |v|^p d\Omega)^{1/p} < \infty\}$ denote the space of all $p$-power integrable functions on $\Omega$ equipped with the usual $L^p$-norm $\|\cdot\|_{L^p(\Omega)}$. We use $(\cdot,\cdot)$ to denote the usual inner product. For any $\zeta = [\zeta_1, \zeta_2] \in (0,1]^2$, we let

$$ H^\zeta(\mathbb{R}^2) := \left\{v : v,\ {}_{-\infty}D_x^{\zeta_1}v \text{and} {}_{-\infty}D_y^{\zeta_2}v \in L^2(\mathbb{R}^2)\right\}. $$

On $H^\zeta(\mathbb{R}^2)$ we introduce an energy norm $\|\cdot\|_\zeta$ such that for any $v \in H^\zeta(\mathbb{R}^2)$,

$$ \|v\|_\zeta^2 = \|v\|_{L^2(\mathbb{R}^2)}^2 + \|{}_{-\infty}D_x^{\zeta_1}v\|_{L^2(\mathbb{R}^2)}^2 + \left\|{}_{-\infty}D_y^{\zeta_2}v\right\|_{L^2(\mathbb{R}^2)}^2. \tag{10} $$

It has been shown in [12] that $H^\zeta(\mathbb{R}^2)$ equipped with $\|\cdot\|_\zeta$ is a Sobolev space.

Similarly to $H^\zeta(\mathbb{R}^2)$, we also define the Sobolev space of functions having a support on $\Omega = (x_{\min}, x_{\max}) \times (y_{\min}, y_{\max})$ given by

$$ H_0^\zeta(\Omega) = \{v : v \in H^\zeta(\Omega), v|_{\partial\Omega} = 0\} $$

with the energy norm defined in (10) (with $\mathbb{R}^2$ replaced with $\Omega$), where ${}_{x_{\min}}D_x^{\zeta_1}u$ and ${}_{y_{\min}}D_y^{\zeta_2}u$ are defined in (8) with $x_0$ and $y_0$ replaced with $x_{\min}$ and $y_{\min}$ respectively and $\partial\Omega$ denotes the boundary of $\Omega$. In what follows, we also use $\langle\cdot,\cdot\rangle$ to denote the duality pairing between $H_0^\zeta(\Omega)$ and its dual space $H_0^{-\zeta}(\Omega)$ defined by $\langle v, w\rangle = \int_\Omega vwd\Omega$ for $v \in H_0^\zeta(\Omega)$ and $w \in H_0^{-\zeta}(\Omega)$.

We first rewrite the operator in (1a) as the following conservative form:

$$ \mathcal{L}U = -U_t - \nabla\cdot\ (-aU + B\nabla^{(\zeta)}U) + rU, $$

where

$$ a = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, \quad \text{and} \quad B = \begin{pmatrix} b_1 & 0 \\ 0 & b_2 \end{pmatrix}. $$

Letting $\gamma = [\alpha, \beta]$ and $\zeta = [\alpha-1, \beta-1]$, we define

$$ \nabla^{(\zeta)}U = \left[\frac{\partial^{\alpha-1}U}{\partial x^{\alpha-1}}, \frac{\partial^{\beta-1}U}{\partial y^{\beta-1}}\right]^T. $$

Let $U_0 \in H^2(\Omega)$ be a known function satisfying the boundary conditions (2c). (For example, $U_0$ can be the solution of a bi-harmonic equation satisfying (2c) and the homogeneous Neumann boundary condition.) Then we introduce a new function

$$ u(x,y,t) = U_0(x,y) - U(x,y,t). \tag{11} $$

DISTRIBUTION A. Approved for public release: distribution unlimited.

Taking $\mathcal{L}U_0$ away from both sides of (2a) and using (11), we have

$$\begin{cases} \mathcal{L}u \le f, \\ u \le u^*, \\ (\mathcal{L}u - f)\cdot(u - u^*) = 0 \end{cases} \tag{12a}$$

for feasible $t$ and $(x, y)$ with the boundary and terminal conditions:

$$u(x, y, t)|_{\partial\Omega} = 0, \quad u(x, y, T) = u^*(x, y), \tag{12b}$$

where $f(x, y) = \mathcal{L}U_0(x, y)$ and $u^*(x, y) = U_0(x, y) - U^*(x, y)$.

We now define

$$\mathcal{K} = \left\{ v(t) : v(t) \in H_0^{\gamma/2}(\Omega), v(t) \le u^*(t) \quad \text{almost everywhere in } (0, T) \right\}.$$

It is easy to verify $\mathcal{K}$ is a convex and closed subset of $H_0^{\gamma/2}(\Omega)$. Using this convex set, we pose the following problem:

**Problem 2.1.** Find $u \in \mathcal{K}$, such that, for all $v \in \mathcal{K}$,

$$\left\langle -\frac{\partial u}{\partial t}, v - u \right\rangle + A(u, v - u) \ge (f, v - u), \tag{13}$$

almost everywhere (*a.e.*) in $(0, T)$, satisfying the boundary and terminal condition (12b), where $A(\cdot, \cdot)$ is a bilinear form defined by:

$$A(u, v) = \langle \nabla u, av \rangle + \langle B\nabla^{(\zeta)} u, \nabla v \rangle + r(u, v), \quad u, v \in H_0^{\gamma/2}(\Omega). \tag{14}$$

it can be easily shown that Problem 2.1 is the variational form of (12a).

In [7] (also [12]), we have proved the following lemma.

**Lemma 2.2.** *For* $\mathcal{A}(u, v) = a\langle \frac{\partial u}{\partial x}, v \rangle + b\langle {}_{x_{\min}}D_x^{\alpha-1} u, \frac{\partial v}{\partial x} \rangle + r(u, v)$, *there exist positive constants* $C_1$ *and* $C_2$, *such that for any* $v, w \in H_0^{\alpha/2}(I)$, $\alpha \in (1, 2)$.

$$\mathcal{A}(v, v) \ge C_1 \|v\|_{\alpha/2}^2, \tag{15}$$

$$\mathcal{A}(v, w) \le C_2 \|v\|_{\alpha/2} \|w\|_{\alpha/2}, \tag{16}$$

*for all* $t \in (0, T)$ *a.e.*

Using this lemma, we can derive the following lemma.

**Lemma 2.3.** *There exist two positive constants* $C_1^*$ *and* $C_2^*$, *such that for any* $v, w \in H_0^{\gamma/2}(\Omega)$,

$$A(v, v) \ge C_1^* \|v\|_{\gamma/2}^2, \tag{17}$$

$$A(v, w) \le C_2^* \|v\|_{\gamma/2} \|w\|_{\gamma/2}, \tag{18}$$

*for* $t \in (0, T)$ *a.e.*

**Proof.** Let $C$ be a generic positive constant. Using (15) and Cauchy–Schwarz inequality, we have, for $u, v \in H_0^{\gamma/2}(\Omega)$,

$$\begin{aligned} A(v, v) &= a_1 \left\langle \frac{\partial v}{\partial x}, v \right\rangle + b_1 \left\langle {}_{x_{\min}}D_x^{\alpha-1} v, \frac{\partial v}{\partial x} \right\rangle + a_2 \left\langle \frac{\partial v}{\partial y}, v \right\rangle + b_2 \left\langle {}_{y_{\min}}D_y^{\beta-1} v, \frac{\partial v}{\partial y} \right\rangle + r(v, v) \\ &\ge C_1 \|v\|_{\alpha/2}^2 + C_2 \|v\|_{\beta/2}^2 \\ &\ge C\left( \|v\|_{\alpha/2} + \|v\|_{\beta/2} \right)^2 \\ &\ge C\|v\|_{\gamma/2}^2. \end{aligned}$$

Similarly, using (16) and Cauchy–Schwarz inequality, we can have

$$\begin{aligned} A(v, w) &= a_1 \left\langle \frac{\partial v}{\partial x}, w \right\rangle + b_1 \left\langle {}_{x_{\min}}D_x^{\alpha-1} v, \frac{\partial w}{\partial x} \right\rangle + a_2 \left\langle \frac{\partial v}{\partial y}, w \right\rangle + b_2 \left\langle {}_{y_{\min}}D_y^{\beta-1} v, \frac{\partial w}{\partial y} \right\rangle + r(v, w) \\ &\le C_1 \|v\|_{\alpha/2} \|w\|_{\alpha/2} + C_2 \|v\|_{\beta/2} \|w\|_{\beta/2} \\ &\le C\left( \|v\|_{\alpha/2}^2 + \|v\|_{\beta/2}^2 \right)^{1/2} \left( \|w\|_{\alpha/2}^2 + \|w\|_{\beta/2}^2 \right)^{1/2} \\ &= C\|v\|_{\gamma/2}^2 \|w\|_{\gamma/2}^2. \end{aligned}$$

□

Using Lemmas 2.2 and 2.3, we are able to prove the following theorem.

DISTRIBUTION A. Approved for public release: distribution unlimited.

**Theorem 2.4.** *There exists a unique solution to Problem 2.1.*

This theorem is just a consequence of Lemma 2.3 and Theorem 1.33 in [14], in which the unique solvability for an abstract variational inequality problem is established. Thus, the proof is omitted here.

To conclude this section, we comment that the transformation (11) is necessary only for theoretical discussions. It is not necessary to use (11) in practical computations.

## 3. Penalty method and convergence

Penalty methods have been used successfully for solving conventional constrained optimization problems. In this section we will propose such a penalty method for (12a) and (12b). We then establish a convergence theory for the penalty method. The penalized equation to solve American-style option pricing problem is given below:

$$\mathcal{L}u_\lambda(x,y,t) + \lambda[u_\lambda(x,y,t) - u^*(x,y)]_+^{1/k} = f(x,y), \qquad (x,y,t) \in \Omega \times (0,T) \tag{19a}$$

satisfying the following boundary and terminal conditions:

$$u_\lambda(x,y,t)|_{\partial\Omega} = 0, \quad u_\lambda(x,y,T) = u^*(x,y), \tag{19b}$$

where $\lambda > 1$ and $k > 0$ are penalty parameters. The variational form of (19) is as follows.

**Problem 3.1.** Find $u_\lambda(t) \in H_0^{\gamma/2}(\Omega)$ satisfying the initial condition in (19b), such that, for all $v \in H_0^{\gamma/2}(\Omega)$,

$$\left\langle -\frac{\partial u_\lambda(t)}{\partial t}, v \right\rangle + A(u_\lambda(t), v) + \left(\lambda[u_\lambda(t) - u^*]_+^{1/k}, v\right) = (f, v) \tag{20}$$

for $t \in (0, T)$ a.e., where $A(\cdot,\cdot)$ is a bilinear form defined in (14).

**Theorem 3.2.** *Problem 3.1 has a unique solution.*

**Proof.** To prove this theorem, it suffices to show that the nonlinear operator on the LHS of (20) is strongly monotone and continuous. Since the linear part $A$ of the LHS of (20) is coercive by (17) and the nonlinear penalty term in (20) is clearly monotone, the operator is strongly monotone.

From (18) we see that $A(u_\lambda, v)$ is Lipschitz continuous in both $u_\lambda$ and $v$. Also, it is obvious that the nonlinear term is continuous in both $u_\lambda$ and $v$. Therefore, Problem 3.1 is uniquely solvable by the standard result in [14, p. 37]. For a more rigorous proof of this theorem, we refer to Theorem 3.2 of [8]. □

We now show that the solution to Problem 3.1 converges to that of (12a) as the penalty parameters $\lambda$ or/and $k \to \infty$ in a proper norm. Before further discussion, it is necessary to introduce the usual Hilbert space in space and time given by

$$L^p(0,T;H(\Omega)) := \{v(\cdot,\cdot;t) : v(\cdot,\cdot;t) \in H(\Omega) \text{ a.e. in } (0,T);\ ||v(\cdot,\cdot;t)||_{H(\Omega)} \in L^p((0,T))\}$$

with the norm

$$\|v(\cdot,\cdot;t)\|_{L^p(0,T;H(\Omega))} = \left(\int_0^T \|v(\cdot,\cdot;t)\|_{H(\Omega)}^p dt\right)^{1/p},$$

where $H(\Omega)$ denotes a Hilbert space on $\Omega$ with the norm $\|\cdot\|_{H(\Omega)}$. Using this space we present the following lemma.

**Lemma 3.3.** *Let $u_\lambda$ be the solution to Problem 3.1 and assume that $u_\lambda \in L^p(\Omega \times (0,T))$, where $p = 1 + 1/k$. Then there exists a positive constant $C$, independent of $u_\lambda$ and $\lambda$, such that*

$$\|[u_\lambda - u^*]_+\|_{L^p(\Omega\times(0,T))} \le \frac{C}{\lambda^k}, \tag{21}$$

$$\|[u_\lambda - u^*]_+\|_{L^\infty(0,T;L^2(\Omega))} + \|[u_\lambda - u^*]_+\|_{L^2(0,T;H_0^{\gamma/2}(\Omega))} \le \frac{C}{\lambda^{k/2}}. \tag{22}$$

**Proof.** Let $C$ be a generic positive constant, independent of $u_\lambda$ and $\lambda$. To simplify notation, we put $\phi(x,y,t) = [u_\lambda(x,y,t) - u^*(x,y)]_+$. It is easy to see that $\phi(\cdot,\cdot,t) \in H_0^{\gamma/2}(\Omega)$ for $t \in (0,T)$ *a.e.* Thus, setting $v = \phi$ in (20), we have

$$\left\langle -\frac{\partial u_\lambda}{\partial t}, \phi \right\rangle + A(u_\lambda, \phi) + \lambda\left(\phi^{1/k}, \phi\right) = (f, \phi) \qquad \text{a.e. in } (0,T).$$

Taking $-\left\langle \frac{\partial u^*}{\partial t}, \phi \right\rangle + A(u^*, \phi)$ away from both sides of the above equality gives

$$\left\langle -\frac{\partial (u_\lambda - u^*)}{\partial t}, \phi \right\rangle + A(u_\lambda - u^*, \phi) + \lambda\left(\phi^{1/k}, \phi\right) = (f, \phi) + \left\langle \frac{\partial u^*}{\partial t}, \phi \right\rangle - A(u^*, \phi),$$

DISTRIBUTION A. Approved for public release: distribution unlimited.

or

$$\left\langle -\frac{\partial \phi}{\partial t}, \phi \right\rangle + A(\phi, \phi) + \lambda\left(\phi^{1/k}, \phi\right) = (f, \phi) - A(u^*, \phi), \tag{23}$$

since $\phi = 0$ when $u_\lambda - u^* < 0$ and $\frac{\partial u^*}{\partial t} = 0$.

Note $\phi(x, y, T) = [u_\lambda(x, y, T) - u^*(x, y)]_+ = 0$ by (19b). Integrating by parts gives

$$\int_t^T \left\langle -\frac{\partial \phi(t)}{\partial t}, \phi(\tau) \right\rangle d\tau = (\phi(t), \phi(t)) - \int_t^T \left\langle -\phi(\tau), \frac{\partial \phi(t)}{\partial t} \right\rangle d\tau$$

from which, we get

$$\int_t^T \left\langle -\frac{\partial \phi(t)}{\partial t}, \phi(\tau) \right\rangle d\tau = \frac{1}{2}(\phi(t), \phi(t)). \tag{24}$$

Integrating (23) from $t$ to $T$ and using (24), (15) and Hölder Inequality , we obtain

$$\begin{aligned}
&\frac{1}{2}(\phi(t), \phi(t)) + C\int_t^T \|\phi(\tau)\|_{\gamma/2}^2 d\tau + \lambda \int_t^T \|\phi(\tau)\|_{L^p(\Omega)}^p d\tau \\
&\quad \le \int_t^T (f(\tau), \phi(\tau)) d\tau - \int_t^T A(u^*, \phi(\tau)) d\tau \\
&\quad \le C\left(\int_t^T \|\phi(\tau)\|_{L^p(\Omega)}^p d\tau\right)^{1/p} - \int_t^T A(u^*, \phi(\tau)) d\tau.
\end{aligned} \tag{25}$$

From the definition of $A(\cdot, \cdot)$ in (14), we see that the integrand of the last term in (25) is

$$-A(u^*, \phi(\tau)) = (au^* + B\nabla^{(\zeta)}u^*, \nabla\phi) + r(u^*, \phi).$$

By Green's theorem, we have

$$\begin{aligned}
-\int_t^T (au^*, \nabla\phi) d\tau &= \int_t^T \int_\Omega \nabla \cdot (au^*)\phi(x, y, \tau) dxdyd\tau - \int_t^T \int_{\partial\Omega} (au^* \cdot n)\phi(x, y, \tau) dxdyd\tau \\
&\le C\int_t^T \int_\Omega dxdy\phi(x, y, \tau) d\tau \le C\left(\int_t^T \|\phi(\tau)\|_{L^p(\Omega)}^p d\tau\right)^{1/p},
\end{aligned}$$

because $U^*$ and $\nabla \cdot (au^*)$ are both bounded on $\bar{\Omega}$, where $n$ denotes the unit vector outward-normal to $\partial\Omega$.

Let $\Omega_1 = \{(x, y) \in \Omega : K - w_1e^x - w_2e^y > 0\}$ and $\Omega_2 = \Omega \setminus \bar{\Omega}_1$ such that $U^*(x, y) = 0$ on $\Omega_2$. We also let $\Gamma_0$ be the interface of $\Omega_1$ and $\Omega_2$ so that $\Gamma_0$ has two opposite orientations: $\Gamma_0^+$ which is oriented in the same direction as $\partial\Omega_1$, and $\Gamma_0^-$ which is oriented in the same direction as $\partial\Omega_2$. Since $\phi = 0$ on $\Gamma_0$, we have, using integration by parts,

$$\begin{aligned}
-(B\nabla^{(\zeta)}u^*, \nabla\phi) &= -\int_{\Omega_1} (B\nabla^{(\zeta)}u^*)^T \nabla\phi dxdy - \int_{\Omega_2} (B\nabla^{(\zeta)}u^*)^T \nabla\phi dxdy \\
&= \int_{\Omega_1} \nabla \cdot (B\nabla^{(\zeta)}u^*)\phi dxdy - \int_{\Gamma_0^+} B\nabla^{(\zeta)}u^* \cdot n\phi ds \\
&\quad + \int_{\Omega_2} \nabla \cdot (B\nabla^{(\zeta)}u^*)\phi dxdy - \int_{\Gamma_0^-} B\nabla^{(\zeta)}u^* \cdot n\phi ds \\
&= -\int_{\Gamma_0^+} B(\nabla^{(\zeta)}u_-^* - \nabla^{(\zeta)}u_+^*) \cdot n\phi ds + \sum_{i=1}^2 \int_{\Omega_i} \nabla \cdot (B\nabla^{(\zeta)}u^*)\phi dxdy,
\end{aligned} \tag{26}$$

where $n$ is the unit outward normal direction of the boundary and $\nabla u_-^*$ and $\nabla u_+^*$ denote the value of $\nabla u^*$ on the left and right sides of $\Gamma_0^+$ respectively. Since $u^* = U_0 - U^*$,

$$\nabla^{(\zeta)}u_\pm^* = \nabla^{(\zeta)}U_{0,\pm} - \nabla^{(\zeta)}U_\pm^*.$$

Note that $U_0 \in H^2(\Omega)$, $\nabla^{(\zeta)}U_0$ is continuous in $\Omega$. From this and (3) we have

$$\nabla^{(\zeta)}u_-^* - \nabla^{(\zeta)}u_+^* = \nabla^{(\zeta)}U_+^* - \nabla^{(\zeta)}U_-^* = -(w_1e^x, w_2e^y)^T,$$

since $\nabla^{(\zeta)}U_-^* = 0$. Since $\Gamma_0$ is characterized by $K - w_1e^x - w_2e^y = 0$, the unit vector outward-normal to $\Gamma_0^+$ is

$$n = \frac{\nabla(K - w_1e^x - w_2e^y)}{\|\nabla(K - w_1e^x - w_2e^y)\|} = \frac{(-w_1e^x, -w_2e^y)^T}{(w_1^2e^{2x} + w_2^2e^{2y})^{1/2}}.$$

Based on the above results, (26) has the following upper bound:

DISTRIBUTION A. Approved for public release: distribution unlimited.

$$-(B\nabla^{(\zeta)}u^*, \nabla\phi) \le -\int_{\Gamma_0^+} \frac{(w_1e^x, w_2e^y)^T B(w_1e^x, w_2e^y)^T}{(w_1^2e^{2x} + w_2^2e^{2y})^{1/2}}\phi ds + \sum_{i=1}^{2}\int_{\Omega_i} \nabla\cdot(B\nabla^{(\zeta)}u^*)\phi dxdy$$
$$\le C\int_\Omega \phi dxdy,$$

since $B$ is positive-definite, $\phi$ is non-negative and $\nabla\cdot(B\nabla^{(\zeta)}u^*)$ is bounded above on both $\Omega_1$ and $\Omega_2$ from its definition. Therefore, replacing the last term in (25) by the above upper bound gives

$$\frac{1}{2}(\phi(t),\phi(t)) + C\int_t^T \|\phi(\tau)\|_{\gamma/2}^2 d\tau + \lambda\int_t^T \|\phi(\tau)\|_{L^p(\Omega)}^p d\tau \le C\left(\int_t^T \|\phi(\tau)\|_{L^p(\Omega)}^p d\tau\right)^{1/p} \tag{27}$$

for all $t \in (0, T)$ a.e. This implies that

$$\lambda\int_t^T \|\phi(\tau)\|_{L^p(\Omega)}^p d\tau \le C\left(\int_t^T \|\phi(\tau)\|_{L^p(\Omega)}^p d\tau\right)^{1/p},$$

and so

$$\left(\int_t^T \|\phi(\tau)\|_{L^p(\Omega)}^p d\tau\right)^{1-1/p} \le C\lambda^{-1}.$$

From the choice of $p$ we see that $1 - 1/p = 1/(kp)$. Thus, from the above estimate we have

$$\left(\int_t^T \|\phi(\tau)\|_{L^p(\Omega)}^p d\tau\right)^{1/p} \le C\lambda^{-k}.$$

This is (21). Combining (27) and the above estimate yields

$$\frac{1}{2}(\phi(t),\phi(t)) + \int_t^T \|\phi(\tau)\|_{\gamma/2}^2 d\tau \le \frac{C}{\lambda^k}$$

for any feasible $t$. Finally, noting that $t$ is arbitrary, the above inequality implies (22). □

Using Lemma 3.3, we are able to prove the following theorem.

**Theorem 3.4.** *Let $u$ and $u_\lambda$ be the solutions to Problems 2.1 and 3.1, respectively. If $\frac{\partial u}{\partial t} \in L^{1+k}(\Omega\times(0,T))$, then there exists a constant $C > 0$, independent of $\lambda$, such that*

$$\|u_\lambda - u\|_{L^\infty(0,T;L^2(\Omega))} + \|u_\lambda - u\|_{L^2(0,T;H_0^{\gamma/2}(\Omega))} \le \frac{C}{\lambda^{k/2}}, \tag{28}$$

*where $\lambda$ and $k$ are the parameters used in (19a).*

**Proof.** Following the notation used in the proof of Lemma 3.3, we decompose $u - u_\lambda$ as

$$u - u_\lambda = u - u^* + [u_\lambda - u^*]_- - [u_\lambda - u^*]_+ =: R_\lambda - \phi, \tag{29}$$

where $[z]_- = -\min\{z, 0\}$ for any $z$ and

$$R_\lambda = u - u^* + [u_\lambda - u^*]_-. \tag{30}$$

Let us first consider $R_\lambda$. Setting $v = u - R_\lambda$ in (13) and $v = R_\lambda$ in (20) gives

$$\left\langle -\frac{\partial u}{\partial t}, -R_\lambda\right\rangle + A(u, -R_\lambda) \ge (f, -R_\lambda),$$
$$\left\langle -\frac{\partial u_\lambda}{\partial t,}, R_\lambda\right\rangle + A(u_\lambda, R_\lambda) + \lambda\left(\phi^{1/k}, R_\lambda\right) = (f, R_\lambda).$$

Adding up the above inequality and equality, we have

$$\left\langle -\frac{\partial(u_\lambda - u)}{\partial t}, R_\lambda\right\rangle + A(u_\lambda - u, R_\lambda) + \lambda\left(\phi^{1/k}, R_\lambda\right) \ge 0. \tag{31}$$

From their definitions, it is easy to see

$$\phi^{1/k}[u_\lambda - u^*]_- = [u_\lambda - u^*]_+^{1/k}[u_\lambda - u^*]_- \equiv 0. \tag{32}$$

Thus, using the above relationship and (30), we have

$$\left(\phi^{1/k}, R_\lambda\right) = \left(\phi^{1/k}, u - u^* + [u_\lambda - u^*]_-\right) = \left(\phi^{1/k}, u - u^*\right) \le 0,$$

DISTRIBUTION A. Approved for public release: distribution unlimited.

since $\phi \geq 0$ and $u - u^* \leq 0$ by (12a). Therefore, (31) reduces to

$$\left\langle -\frac{\partial (u-u_\lambda)}{\partial t}, R_\lambda \right\rangle + A(u-u_\lambda, R_\lambda) \leq 0.$$

Using (29), it is easy to see that the above inequality can be rewritten as

$$\left\langle -\frac{\partial R_\lambda}{\partial t}, R_\lambda \right\rangle + A(R_\lambda, R_\lambda) \leq \left\langle -\frac{\partial \phi}{\partial t}, R_\lambda \right\rangle + A(\phi, R_\lambda).$$

From (30) we see that $R_\lambda(x,T) = 0$. Thus, integrating both sides of the above estimate from $t$ to $T$ and using the same argument as for (24), Cauchy-Schwarz inequality and (18), we have

$$\begin{aligned}
&\frac{1}{2}(R_\lambda(t), R_\lambda(t)) + \int_t^T A(R_\lambda(\tau), R_\lambda(\tau))d\tau \\
&\leq \int_t^T \left\langle -\frac{\partial \phi(\tau)}{\partial \tau}, R_\lambda(\tau) \right\rangle d\tau + \int_t^T A(\phi(\tau), R_\lambda(\tau))d\tau \\
&\leq (\phi(t), R_\lambda(t)) + \int_t^T \left\langle \phi(\tau), \frac{\partial R_\lambda(\tau)}{\partial \tau} \right\rangle d\tau + \int_t^T A(\phi(\tau), R_\lambda(\tau))d\tau \\
&\leq \|\phi\|_{L^\infty(0,T;L^2(\Omega))} \|R_\lambda\|_{L^\infty(0,T;L^2(\Omega))} + C\|\phi\|_{L^2(0,T;H_0^{\gamma/2}(\Omega))} \|R_\lambda\|_{L^2(0,T;H_0^{\gamma/2}(\Omega))} \\
&\quad + \int_t^T \left\langle \phi(\tau), \frac{\partial R_\lambda(\tau)}{\partial \tau} \right\rangle d\tau,
\end{aligned} \tag{33}$$

for all $t \in (0, T)$. Using (32), (30), and (21), we estimate the last term in (33) as follows:

$$\int_t^T \left\langle \phi(\tau), \frac{\partial R_\lambda(\tau)}{\partial \tau} \right\rangle d\tau = \int_t^T \left\langle \phi(\tau), \frac{\partial u(\tau)}{\partial \tau} \right\rangle d\tau \leq C\|\phi\|_{L^p(\Omega\times(0,T))} \left\| \frac{\partial u}{\partial t} \right\|_{L^q(\Omega\times(0,T))} \leq \frac{C}{\lambda^k},$$

where $p = 1 + 1/k$ and $q = 1 + k$. Substituting the above upper bound into (33) and using (16), (15) and (22), we obtain

$$\begin{aligned}
&\left(\|R_\lambda\|_{L^\infty(0,T;L^2(\Omega))} + \|R_\lambda\|_{L^2(0,T;H^{\gamma/2}(\Omega))}\right)^2 \\
&\leq C\left(\frac{1}{2}\|R_\lambda\|^2_{L^\infty(0,T;L^2(\Omega))} + \|R_\lambda\|^2_{L^2(0,T;H^{\gamma/2}(\Omega))}\right) \\
&\leq C\left[\|\phi\|_{L^\infty(0,T;L^2(\Omega))}\|R_\lambda\|_{L^\infty(0,T;L^2(\Omega))} + \|\phi\|_{L^2(0,T;H_0^{\gamma/2}(\Omega))}\|R_\lambda\|_{L^2(0,T;H_0^{\gamma/2}(\Omega))} + \lambda^{-k}\right] \\
&\leq C\left[\left(\|\phi\|_{L^\infty(0,T;L^2(\Omega))} + \|\phi\|_{L^2(0,T;H^{\gamma/2}(\Omega))}\right)\cdot\left(\|R_\lambda\|_{L^\infty(0,T;L^2(\Omega))} + \|R_\lambda\|_{L^2(0,T;H^{\gamma/2}(\Omega))}\right) + \lambda^{-k}\right] \\
&\leq C\left[\lambda^{-k/2}\left(\|R_\lambda\|_{L^2(0,T;L^2(\Omega))} + \|R_\lambda\|_{L^2(0,T;H^{\gamma/2}(\Omega))}\right) + \lambda^{-k}\right].
\end{aligned}$$

This is of the form $\rho^2 \leq C(\rho\lambda^{-k/2} + \lambda^{-k})$. It is easy to prove that $\rho \leq C\lambda^{-k/2}$ for a generic positive constant $C$, independent of $\lambda$ and $k$. Therefore, we have

$$\|R_\lambda\|_{L^\infty(0,T;L^2(\Omega))} + \|R_\lambda\|_{L^2(0,T;H^{\gamma/2}(\Omega))} \leq C\lambda^{-k/2}. \tag{34}$$

Finally, using the triangular inequality, (29), (22) and (34), we can have

$$\begin{aligned}
\|u-u_\lambda\|_{L^\infty(0,T;L^2(\Omega))} + \|u-u_\lambda\|_{L^2(0,T;H^{\gamma/2}(\Omega))} &\leq \left(\|R_\lambda\|_{L^\infty(0,T;L^2(\Omega))} + \|R_\lambda\|_{L^2(0,T;H^{\gamma/2}(\Omega))}\right) \\
&+ \left(\|\phi\|_{L^\infty(0,T;L^2(\Omega))} + \|\phi\|_{L^2(0,T;H^{\gamma/2}(\Omega))}\right) \leq C\lambda^{-k/2}.
\end{aligned}$$

This is (28). □

## 4. Discretization

Since the penalized fPDE cannot be solved analytically, it needs to be discretized in order to solve it numerically. Various discretization methods are available in the open literature. In this section, we apply the discretization technique developed recently in [7] to the fractional derivatives in (19a). We also use Crank–Nicolson time stepping method to construct the discretization scheme for the penalized equation of (2).

Let the intervals $(x_{\min}, x_{\max})$ and $(y_{\min}, y_{\max})$ be divided into $M_x$ and $M_y$ sub-intervals respectively with mesh nodes

$$x_i = x_{\min} + ih_x, \qquad i = 0, 1, \ldots, M_x; \qquad y_j = y_{\min} + jh_y, \qquad j = 0, 1, \ldots, M_y,$$

where $h_x = (x_{\max} - x_{\min})/M_x$ and $h_y = (y_{\max} - y_{\min})/M_y$. The $\alpha$-th partial derivative defined in (9) can be approximated as follows [7]:

$${}_{x_{\min}}D_x^\alpha U(x_i, y_j, t) \approx \frac{h_x^{-\alpha}}{\Gamma(2-\alpha)} \sum_{k=0}^{i+1} g_k^\alpha U_{i-k+1,j} \tag{35}$$

DISTRIBUTION A. Approved for public release: distribution unlimited.

for any $i \in \{1, 2, \ldots, M_x - 1\}$ and $j \in \{1, 2, \ldots, M_y - 1\}$, where $U_{i-k+1,j} = U(x_{i-k+1}, y_j, t)$. The coefficients $g_k^\alpha$'s are given by

$$g_0^\alpha = \frac{1}{(2-\alpha)(3-\alpha)}, \qquad g_1^\alpha = \frac{2^{3-\alpha} - 4}{(2-\alpha)(3-\alpha)}, \qquad g_2^\alpha = \frac{3^{3-\alpha} - 4 \times 2^{3-\alpha} + 6}{(2-\alpha)(3-\alpha)},$$
$$g_k^\alpha = \frac{1}{(2-\alpha)(3-\alpha)}[(k+1)^{3-\alpha} - 4k^{3-\alpha} + 6(k-1)^{3-\alpha} - 4(k-2)^{3-\alpha} + (k-3)^{3-\alpha}],$$

for $k = 3, 4, \ldots, i+1$. This finite difference scheme has second order accuracy as proved in [7].

For a positive integer $N$, let $(0, T)$ be divided into $N$ sub-intervals with the mesh points $t_n = T - n\Delta t,\ n = 0, 1, \ldots, N$, where $\Delta t = T/N$. Thus $T = t_0 > t_1 > \cdots > t_N = 0$. Using (35) we define the following finite difference operators for the fractional derivatives in (1a):

$$\delta_x^\alpha U_{i,j}^n = \frac{1}{h_x^\alpha \Gamma(2-\alpha)} \sum_{k=0}^{i+1} g_k^\alpha U_{i-k+1,j}^n, \quad \delta_y^\beta U_{i,j}^n = \frac{1}{h_y^\beta \Gamma(2-\beta)} \sum_{k=0}^{j+1} g_k^\beta U_{i,j-k+1}^n, \tag{36a}$$

where $U_{p,q}^n$ denotes an approximation to $U(x_p, y_q, t_n)$ for all feasible $(p, q, n)$. We also define the following central difference approximations to $U_x$ and $U_y$ respectively:

$$\delta_x U_{i,j}^n = \frac{1}{2h_x}(U_{i+1,j}^n - U_{i-1,j}^n), \quad \delta_y U_{i,j}^n = \frac{1}{2h_y}(U_{i,j+1}^n - U_{i,j-1}^n). \tag{36b}$$

Using Crank–Nicolson time stepping method and the finite differences defined in (36a) and (36b), we construct the following discretization scheme for (19a):

$$\frac{U_{i,j}^{n+1} - U_{i,j}^n}{\Delta t} + \frac{1}{2}\left(a_1\delta_x U_{i,j}^{n+1} - b_1\delta_x^\alpha U_{i,j}^{n+1} + a_2\delta_y U_{i,j}^{n+1} - b_2\delta_y^\beta U_{i,j}^{n+1} + rU_{i,j}^{n+1} + d_{ij}^{n+1}\right)$$
$$+\frac{1}{2}\left(a_1\delta_x U_{i,j}^n - b_1\delta_x^\alpha U_{i,j}^n + a_2\delta_x U_{i,j}^n - b_2\delta_y^\beta U_{i,j}^n + rU_{i,j}^n + d_{i,j}^n\right) = 0 \tag{37a}$$

for $i = 1, 2, \ldots, M_x - 1,\ j = 1, 2, \ldots, M_y - 1$ and $n = 1, 2, \ldots, N$ satisfying

$$U_{0,j}^n = g_1(y_j, t_n), \quad U_{i,0}^n = g_2(x_i, t_n), \quad U_{i,j}^0 = U^*(x_i, y_j), \tag{37b}$$

where $d_{i,j}^n := d(U_{i,j}^n) = \lambda[U_{i,j}^n - U_{i,j}^*]_+^{1/k}$ is the penalty term.

Eq. (37a) can be rewritten as the following linear system:

$$\left(\mathbf{I} + \frac{1}{2}\mathbf{M}\right)\mathbf{V}^{n+1} + \frac{1}{2}\mathbf{D}(\mathbf{V}^{n+1}) = \left(\mathbf{I} - \frac{1}{2}\mathbf{M}\right)\mathbf{V}^n - \frac{1}{2}\mathbf{D}(\mathbf{V}^n) + \bar{\mathbf{F}}^n$$

with

$$\mathbf{V}^n = \left(U_{1,1}^n, U_{2,1}^n, \ldots, U_{M_x-1,1}^n, \ldots, U_{M_x-1,M_y-1}^n\right)^T,$$
$$\mathbf{D}(\mathbf{V}^n) = \left(d(U_{1,1}^n), d(U_{2,1}^n), \ldots, d(U_{M_x-1,1}^n), \ldots, d(U_{M_x-1,M_y-1}^n)\right)^T,$$

for $n = 0, 1, \ldots, N-1$, where $\mathbf{I}$ is an $(M_x - 1)(M_y - 1)$-dimensional identity, $\bar{\mathbf{F}}^n$ is an $(M_x - 1) \times (M_y - 1)$ column vector representing the average of the contributions of the boundary conditions at time levels $n$ and $n+1$, and $\mathbf{M}$ is a block matrix containing $(M_y - 1) \times (M_y - 1)$ blocks . The size of each block is $(M_x - 1) \times (M_x - 1)$.

$$\mathbf{M} = \begin{bmatrix} \mathbf{A}+\mathbf{B}_1 & \mathbf{B}_0 & \mathbf{0} & \cdots & \cdots & \cdots & \mathbf{0} \\ \mathbf{B}_2 & \mathbf{A}+\mathbf{B}_1 & \mathbf{B}_0 & \ddots & & & \mathbf{0} \\ \mathbf{B}_3 & \mathbf{B}_2 & \mathbf{A}+\mathbf{B}_1 & \mathbf{B}_0 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{B}_{M_y-3} & & \ddots & \mathbf{B}_2 & \mathbf{A}+\mathbf{B}_1 & \mathbf{B}_0 & \mathbf{0} \\ \mathbf{B}_{M_y-2} & & & \mathbf{B}_3 & \mathbf{B}_2 & \mathbf{A}+\mathbf{B}_1 & \mathbf{B}_0 \\ \mathbf{B}_{M_y-1} & \cdots & \cdots & \cdots & \mathbf{B}_3 & \mathbf{B}_2 & \mathbf{A}+\mathbf{B}_1 \end{bmatrix}_{(M_y-1)\times(M_y-1)}$$

DISTRIBUTION A. Approved for public release: distribution unlimited.

**Table 1**
System and market parameters for the two-asset American option.

| | | | |
|---|---|---|---|
| $\alpha$ , $\beta$ | 1.5 | $r$ | 0.05 |
| $\sigma$ | 0.25 | $K$ | 30 |
| $a_1$, $a_2$ | 0.384 | $b_1$, $b_2$ | 0.884 |
| $x_{\min}$ , $y_{\min}$ | ln 0.1 | $x_{\max}$ , $y_{\max}$ | ln 100 |

**Table 2**
Convergence behavior in $\lambda$.

| $\lambda = 10 \times 2^n$ | | $n = 1$ | $n = 2$ | $n = 3$ | $n = 4$ |
|---|---|---|---|---|---|
| $k = 1$ | Error | 1.3229 | 0.6857 | 0.3493 | 0.1763 |
| | $\log_2$Ratio | | 0.9480 | 0.9731 | 0.9863 |
| $k = 2$ | Error | 1.6211 | 0.4752 | 0.1246 | 0.0315 |
| | $\log_2$Ratio | | 1.7704 | 1.9307 | 1.9825 |
| $k = 3$ | Error | 1.8691 | 0.3343 | 0.0447 | 0.006 |
| | $\log_2$Ratio | | 2.4833 | 2.9036 | 3.0010 |

In the above expression, $\mathbf{A}$ is the discretization matrix on $x$ direction in (35)

$$\mathbf{A}_{ij} = \begin{cases} \mu_x g_0^\alpha + \eta_x, & j = i+1 \\ \mu_x g_1^\alpha + \frac{r}{2}\Delta t, & j = i \\ \mu_x g_2^\alpha - \eta_x, & j = i-1 \\ \mu_x g_l^\alpha, & j = i-l+1, l = 3, \ldots, i \\ 0, & \text{otherwise}, \end{cases} \qquad \mathbf{B}_j = \begin{cases} (\mu_y g_0^\beta + \eta_y)\mathbf{I}_y, & j = 0 \\ (\mu_y g_1^\beta + \frac{r}{2}\Delta t)\mathbf{I}_y, & j = 1 \\ (\mu_y g_2^\beta - \eta_x)\mathbf{I}_y, & j = 2 \\ (\mu_y g_j^\beta)\mathbf{I}_y, & j = 3, \ldots, M_y, \end{cases}$$

where $\mu_x = -b_1 \frac{\Delta t}{\Gamma(2-\alpha)h_x^\alpha}$, $\eta_x = a_1 \frac{\Delta t}{2h_x}$, $\mu_y = -b_2 \frac{\Delta t}{\Gamma(2-\beta)h_y^\beta}$, and $\eta_y = a_2 \frac{\Delta t}{2h_y}$. Note that the boundary conditions $g_{1j}^n$ and $g_{2i}^n$ (37b) have to be defined by the numerical solutions of two 1D systems.

The nonlinear system (37) can be solved by the following damped Newton's iterative method:

$$\left(\mathbf{I} + \frac{1}{2}\mathbf{M} + \frac{1}{2}J_D\left(w^{l-1}\right)\right)\delta w^{l-1} = \left(\mathbf{I} - \frac{1}{2}\mathbf{M}\right)\mathbf{V}^n - \frac{1}{2}\mathbf{D}(\mathbf{V}^n) + \bar{\mathbf{F}}^n - \left(\mathbf{I} + \frac{1}{2}\mathbf{M}\right)w^{l-1} - \frac{1}{2}\mathbf{D}\left(w^{l-1}\right),$$
$$w^l = w^{l-1} + \kappa\delta w^{l-1}$$

for $l = 1, 2, \ldots$ until a convergence criterion is satisfied with the initial guess $w^0 = \mathbf{V}^n$. $J_D(w)$ denotes the Jacobian matrix of the column vectors $\mathbf{D}(w)$ and $\kappa \in (0, 1]$ denotes a damping parameter. Then we choose $\mathbf{V}^{n+1} = \lim_{l\to\infty} w^l$ for all $n = 0, 1, 2, \ldots, N-1$.

## 5. Numerical experiments

In this section, we present some numerical experimental results to verify the theoretical rate of convergence obtained in Section 3 and the rate of convergence of the discretization scheme in Section 4 to demonstrate the accuracy and usefulness of our numerical method. To achieve this, we use two examples and the first test example is chosen to be the following American basket option pricing problem.

**Example 5.1** (American basket put option pricing)**.** The fractional differential LCP (2) with system and market parameters given in Table 1 and the weights $w_1 = w_2 = 0.5$.

To investigate the convergence rates of the method in both $\lambda$ and $k$, we choose a fixed uniform mesh for the solution domain $(\ln(0.05), \ln(100))^2 \times (0, 1)$ in $(x, y, t)$ with $M_x = M_y = 50$ and $N = 50$. Since the exact solution to this problem is unknown, we use the numerical solution with $\lambda = 10^6$ and $k = 1$ as the reference solution denoted as $V_R$. We solve (19a) on the aforementioned uniform mesh for a sequence of values of $\lambda$ and a fixed value of $k$, and compute approximations of the following continuous norm on the mesh using the reference and numerical solutions $V_R$ and $V_\lambda$:

$$\|V_R - V_\lambda\|_{L^\infty(0,T;L^2(\Omega))} + \|V_R - V_\lambda\|_{L^2(0,T;H^{\gamma/2}(\Omega))}.$$

We also calculate the base-2 logarithm of the ratio of the errors from two consecutive values of $\lambda$ for a fixed $k$ and the results are listed in Table 2. From Theorem 3.4 we see that the ratio of the errors in $V_\lambda$ and $V_{\lambda/2}$ for a given $k$ behaves like $2^{k/2}$. However, from Table 2 we see that the computed ratios behave like $2^k$, indicating that the rate of convergence is of order $\lambda^{-k}$. In fact, it has been proved in [15,26,28], using the fact that all the norms in finite dimensions are equivalent, that the power penalty method for a nonlinear complementarity problems in finite dimensions satisfying a strong monotone condition has the convergence rate $\mathcal{O}(\lambda^{-k})$. However, the convergence rate in finite dimensions is not uniform in

DISTRIBUTION A. Approved for public release: distribution unlimited.

**Table 3**
Convergence behavior in $k$.

| | | $k=1$ | $k=2$ | $k=3$ |
|---|---|---|---|---|
| $\lambda=40$ | Error | 0.6857 | 0.4752 | 0.3343 |
| | $\log_2$Ratio | | 1.4430 | 1.4216 |
| $\lambda=80$ | Error | 0.3493 | 0.1246 | 0.0447 |
| | $\log_2$Ratio | | 2.8025 | 2.7904 |
| $\lambda=160$ | Error | 0.1763 | 0.0315 | 0.0060 |
| | $\log_2$Ratio | | 5.5900 | 5.6529 |

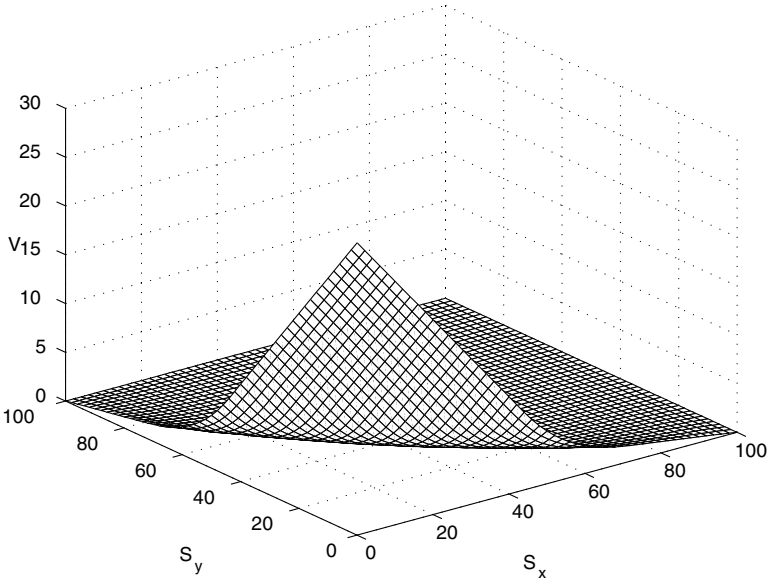



**Fig. 1.** Computed prices of an American Basket option when $\alpha = 1.5$.

the dimensionality. Since norms on an infinite-dimensional space are usually not equivalent, we are unable to achieve the $\mathcal{O}(\lambda^{-k})$-rate of convergence as in finite dimensions.

We now investigate numerically the rate of convergence of the method in $k$ for a fixed $\lambda > 1$. From (28) we see the ratio of the errors in the solutions using $k$ and $k+1$ equals $\mathcal{O}(\lambda^{(k+1)/2}/\lambda^{k/2}) = \mathcal{O}(\lambda^{1/2})$, i.e., the ratio is a constant for any $k$. The computed results for different values of $k$ and $\lambda$ are listed in Table 3, from which we see that the ratios of the errors for any two consecutive values of $k$ are almost constants.

The solution when $\alpha = \beta = 1.5$ is illustrated in Fig. 1. We have also repeated the above numerical experiments for $\alpha = \beta = 1.3, 1.7$ and found that the computed convergence rates are the same as the corresponding ones for $\alpha = \beta = 1.5$, which suggests that the convergence rates of the penalty method do not depend on the fractional order $\alpha$ or $\beta$.

To see the influence of $\alpha$ and $\beta$ on the option price, we solve the test problem for $\alpha = \beta = 1.3, 1.5, 1.7$, and plot the cross-sections at $S_x = S_y$, $0 < S_x < 100$, and $t = 0$ in Fig. 2 of the differences between the numerical solutions, $V_{FBS}$, of the test problem and the numerical solutions of the standard American option $V_{BS}$ (i.e., $\alpha = \beta = 2$). From Fig. 2, we see that the American put option from the fractional model is more valuable than that from the standard model. Also, the value of the option increases as $\alpha$ and $\beta$ decreases. This phenomenon is reasonable as when $\alpha$ and $\beta$ are smaller, the price movement is faster and thus the option premium is higher, similar to the case that the larger the volatility, the higher the option premium.

**Example 5.2** (Fractional advection–diffusion equation)**.** To test the rate of convergence of the discretization scheme we choose the following linear fPDE to which the exact solution is known:

$$u_t + u_x - {}_0D_x^{\alpha}u - {}_0D_y^{\beta}u = f(x,y,t)$$

with boundary and terminal conditions

$$u(x,0,t) = u(x,1,t) = 0, \quad y \in (0,1), \quad t \in (0,1],$$
$$u(0,y,t) = u(1,y,t) = 0, \quad x \in (0,1), \quad t \in (0,1],$$
$$u(x,y,1) = x^3y^4, \quad (x,y) \in (0,1)\times(0,1],$$

where $f(x,y,t) = x^3y^4 + (3x^2y^4 - \frac{\Gamma(4)}{\Gamma(4-\alpha)}x^{3-\alpha}y^4 + 4x^3y^3 - \frac{\Gamma(5)}{\Gamma(5-\beta)}x^3y^{4-\beta} - x^3y^4)t$.

DISTRIBUTION A. Approved for public release: distribution unlimited.

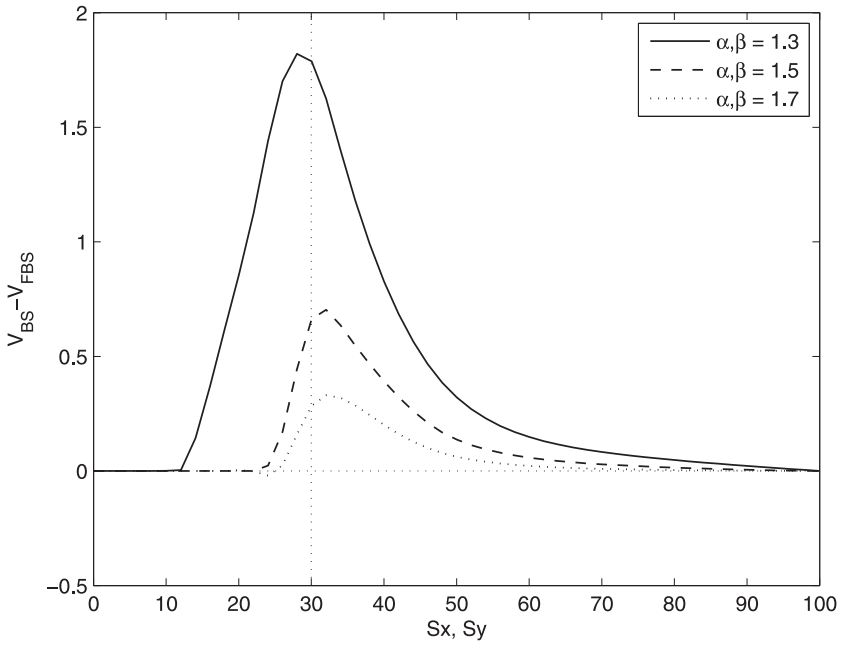



**Fig. 2.** Basket Option comparison for different $\alpha$.

**Table 4**
Computed rates of convergence for Example 5.2.

| $h = \Delta t = \frac{1}{5\times 2^m}$ | $E_m^{GL}$ | $\log_2 \frac{E_m^{GL}}{E_{m+1}^{GL}}$ | $E_m$ | $\log_2 \frac{E_m}{E_{m+1}}$ |
|---|---|---|---|---|
| $m = 0$ | 1.9403e−02 | | 1.7505e−03 | |
| $m = 1$ | 1.4090e−02 | 0.4616 | 5.4035e−04 | 1.6958 |
| $m = 2$ | 8.2916e−03 | 0.7650 | 1.4873e−04 | 1.8612 |
| $m = 3$ | 4.4631e−03 | 0.8936 | 3.8689e−05 | 1.9427 |
| $m = 4$ | 2.3158e−03 | 0.9466 | 9.8882e−06 | 1.9681 |
| $m = 5$ | 1.1801e−03 | 0.9726 | 2.5081e−06 | 1.9791 |

We choose $\alpha = \beta = 1.5$ and the exact solution to the above problem is $u(x,t) = x^3y^4t$. This problem is solved using a sequence of uniform meshes with mesh sizes $h_x = h_y = h = \Delta t = \frac{1}{5}\times 2^{-m}$ for $m = 0, 1, \ldots, 5$. For each $m$, the following discrete maximum norm is computed:

$$E_m = \max_{0<n<N-1,\ 1<i<M_x-1,\ 1<j<M_y-1} \left|u(x_i, y_j, t_n) - U_{ij}^n\right|,$$

where $(U_{ij}^n)$ denotes the numerical solution by the discretization scheme. These computed errors, along with computed rates of convergence $\log_2(E_{m+1}/E_m)$, for $k = 0, 1, \ldots, 4$, are listed in Table 4 from which we see that the rates of convergence of our method are of order $\mathcal{O}(\Delta t^2 + h_x^2 + h_y^2)$, while a lengthy mathematical proof of this upper error bound can be found in [9]. For comparison, we have also solved this 2D problem using the combination of the Crank–Nicolson time-stepping scheme and Grünwald–Letnikov method in [23] which is a popular method for fPDEs. The computed errors $E_m^{GL}$'s and the rates of convergence for the GL method are also listed in Table 4, from which it is clear that GL method is only 1st-order accurate, and our method has a 2nd-order convergence rate.

Finally, we comment that the coefficient matrix **M** of the discretized system in Section 4 is dense and thus the computational costs for solving the discretized system is usually high, particularly in 2 spatial dimensions. Theoretically, it is known that the computational cost for solving the system using the LU decomposition is of the order $\mathcal{O}((M_x \times M_y)^3)$. The development of efficient algorithms for (37) such as conjugate gradient based and ADI algorithms is a future topic and challenge for us, though it is beyond our current discussion. Also, a comparison of numerical performances of a penalty method similar to the current one with the augmented Lagrangian method for solving conventional American option pricing problems has been given in [32]. Thus we refer readers to this comparison study.

## 6. Conclusion

In this paper, we proposed and analyzed a power penalty method a 2-dimensional fractional differential linear complementarity problem for pricing American options on two independent assets. We proved that the solution from the penalty

DISTRIBUTION A. Approved for public release: distribution unlimited.

method converges to that of the linear complementarity problem at the rate of $\mathcal{O}(\lambda^{-k/2})$. A 2nd-order accurate discretization scheme has also been developed for solving the nonlinear fractional partial differential equation arising from the penalty approach. Numerical experiments were performed to verify the theoretical rates of convergence and demonstrate that numerical method produces financially meaningful results.

## Acknowledgments

This work was partially supported by the AOARD Project #15IOA095 from the US Air Force Office of Scientific Research.

## References

[1] A. Almendral, C.W. Oosterlee, Numerical valuation of options with jumps in the underlying, Appl. Numer. Math. 53 (2005) 1–18.
[2] A. Anderson, J. Andreasen, Jump diffusion process: volatility smile fitting and numerical methods for option pricing, Rev. Deriv. Res. 4 (2000) 231–262.
[3] A. Bensoussan, J.L. Lions, Applications of Variational Inequalities in Stochastic Control, North-Holland, Amsterdam, New York, Oxford, 1978.
[4] F. Black, M. Scholes, The pricing of options and corporate liabilities, J. Polit. Econ. (1973) 637–654.
[5] P. Carr, H. Geman, D. Madan, M. Yor, The fine structure of asset returns: an empirical investigation, J. Bus. 75 (2) (2002) 305–332.
[6] A. Cartea, D. del Castillo-Negrete, Fractional diffusion models of option prices in markets with jumps, Phys. A: Stat. Mech. Appl. 374 (2) (2007) 749–763.
[7] W. Chen, S. Wang, A finite difference method for pricing European and American options under a geometric Lévy process, J. Ind. Manag. Optim. 11 (1) (2015) 241–264.
[8] W. Chen, S. Wang, A penalty method for a fractional order parabolic variational inequality governing American put option valuation, Comput. Math. Appl. 67 (1) (2014) 77–90.
[9] W. Chen, S. Wang, A second order finite difference method for a 2D fractional Black–Scholes equation, in Lecture Notes in Computer Science, Vol. 10187, Springer, in press.
[10] S.S. Clift, P.A. Forsyth, Numerical solution of two asset jump diffusion models for option valuation, Appl. Numer. Math. 58 (6) (2008) 743–782.
[11] S. De Cosmis, R. De Leone, The use of grossone in mathematical programming and operations research, Appl. Math. Comput. 218 (16) (2012) 8029–8038.
[12] V.J. Ervin, J.P. Roop, Variational formulation for the stationary fractional advection dispersion equation, Numer. Methods Partial Differ. Equ. 22 (3) (2006) 558–576.
[13] Y. d'Halluin, P.A. Forsyth, K.R. Vetzal, Robust numerical methods for contingent claims under jump diffusion processes, IMA J. Numer. Anal. 25 (2005) 87–112.
[14] J. Haslinger, M. Miettinen, Finite Element Method for Hemivariational Inequalities, Kluwer Academic Publisher, Dordrecht, Boston, London, 1999.
[15] C.C. Huang, S. Wang, A power penalty approach to a nonlinear complementary problem, Oper. Res. Lett. 38 (2010) 72–76.
[16] C.S. Huang, C.H. Hung, S. Wang, A fitted finite volume method for the valuation of options on assets with stochastic volatilities, Computing 77 (2006) 297–320.
[17] C.S. Huang, C.H. Hung, S. Wang, On convergence of a fitted finite-volume method for the valuation of options on assets with stochastic volatilities, IMA J. Numer. Anal. 30 (2010) 1101–1120.
[18] W. Li, S. Wang, Pricing American options under proportional transaction costs using a penalty approach and a finite difference scheme, J. Ind. Manag. Optim. 9 (2013) 365–389.
[19] W. Li, S. Wang, A numerical method for pricing European options with proportional transaction costs, J. Global Optim. 60 (2014) 59–78.
[20] R.C. Merton, Theory of rational option pricing, Bell J. Econ. Manag. Sci. 4 (1973) 141–183.
[21] K.S. Miller, B. Ross, An Introduction to the Fractional Calculus and Fractional Differential Equations, John Wiley & Sons, New York, NY, USA, 1993.
[22] B.F. Nielsen, O. Skavhaug, A. Tveito, Penalty and front-fixing methods for the numerical solution of American option problems, J. Comp. Fin. 5 (2001) 69–97.
[23] K.B. Oldham, J. Spanier, The fractional calculus, 1974, pp. 76–81.
[24] Y.D. Sergeyev, Arithmetic of Infinity, Edizioni Orizzonti Meridionali, Cosenza, 2003.
[25] A.M. Rubinov, X.Q. Yang, Lagrange-type Functions in Constrained Non-Convex Optimization, Kluwer Academic Publishers, 2003.
[26] S. Wang, X.Q. Yang, A power penalty method for linear complementarity problems, Oper. Res. Lett. 36 (2008) 211–214.
[27] S. Wang, X.Q. Yang, K.L. Teo, Power penalty method for a linear complementarity problem arising from American option valuation, J. Optim. Theory Appl. 129 (2) (2006) 227–254.
[28] S. Wang, A penalty method for a finite-dimensional obstacle problem with derivative constraints, Optim. Lett. 8 (2014) 1799–1811.
[29] P. Wilmott, J. Dewynne, S. Howison, Option Pricing: Mathematical Models and Computation, Oxford Financial Press, 1993.
[30] K. Zhang, S. Wang, Pricing options under jump diffusion processes with fitted finite volume method, Appl. Math. Comput. 201 (2008) 398–413.
[31] K. Zhang, S. Wang, A computational scheme for options under jump diffusion processes, Int. J. Numer. Anal. Model. 6 (2009) 110–123.
[32] K. Zhang, X.Q. Yang, S. Wang, K.L. Teo, Numerical performances of penalty method for American option pricing, Optim. Methods Softw. 25 (2010) 737–752.
[33] K. Zhang, S. Wang, Convergence property of an interior penalty approach to pricing American option, J. Ind. Manag. Optim. 7 (2011) 435–447.
[34] K. Zhang, S. Wang, Pricing American bond options using a penalty method, Automatica 48 (2012) 472–479.
[35] Y.Y. Zhou, S. Wang, X.Q. Yang, A penalty approximation method for a semilinear parabolic double obstacle problem, J. Global Optim. 60 (2014) 531–550.
[36] R. Zvan, P.A. Forsyth, K.R. Vetzal, Penalty methods for American options with stochastic volatility, Comput. Appl. Math. 91 (1998) 199–218.

DISTRIBUTION A. Approved for public release: distribution unlimited.

Computers and Mathematics with Applications 73 (2017) 2454–2469

Contents lists available at ScienceDirect

Computers and Mathematics with Applications

journal homepage: www.elsevier.com/locate/camwa

# Pricing European options with proportional transaction costs and stochastic volatility using a penalty approach and a finite volume scheme

Wen Li, Song Wang *

*Department of Mathematics & Statistics, Curtin University, GPO Box U1987, Perth WA6845, Australia*

ARTICLE INFO

*Article history:*
Received 3 October 2016
Accepted 26 March 2017
Available online 29 April 2017

*Keywords:*
Hamilton–Jacobi–Bellman equation
Financial option valuation
Finite volume method
Penalty method
Convergence

ABSTRACT

In this paper we propose a combination of a penalty method and a finite volume scheme for a four-dimensional time-dependent Hamilton–Jacobi–Bellman (HJB) equation arising from pricing European options with proportional transaction costs and stochastic volatility. The HJB equation is first approximated by a nonlinear partial differential equation containing penalty terms. A finite volume method along with an upwind technique is then developed for the spatial discretization of the nonlinear penalty equation. We show that the coefficient matrix of the discretized system is an $M$-matrix. An iterative method is proposed for solving the nonlinear algebraic system and a convergence theory is established for the iterative method. Numerical experiments are performed using a non-trivial model pricing problem and the numerical results demonstrate the usefulness of the proposed method.

© 2017 Elsevier Ltd. All rights reserved.

## 1. Introduction

Valuation of options is one of the most important problems in financial engineering. For over four decades, practitioners and academic researchers in finance, economics and mathematics have engaged in the study of option pricing. Various option pricing approaches have been developed (see, for example, [1–9]). One of the methods is the utility based option pricing approach which has been widely used for valuing European and American options when the trading of the underlying stocks incurs proportional transaction costs [3,6,10–18]. Recently, Caflisch et al. [19] and Cosso [20] applied this approach to pricing European options and American options respectively under proportional transaction costs and stochastic volatility. More specifically, in [19] the authors assumed that the underlying stock price follows a geometric Brownian motion and the associated volatility evolves according to a stochastic process of the Ornstein–Uhlenbeck type. By following the utility maximization procedure proposed in [3], they derived a set of non-linear HJB equations governing European option prices. They also obtained an asymptotic expression for the European option price in the limit of small transaction costs and fast mean reversion volatility under the assumption of an exponential utility function. In [20], the authors considered American option pricing with proportional transaction costs and stochastic volatility. They assumed that the stochastic volatility follows the Cox–Ingersoll–Ross (CIR) process. They also showed that computing the price of an American option involves solving a singular stochastic optimal control problem and proved the existence and uniqueness of the viscosity solution to the associated HJB equation. Moreover, they solved the HJB equations using the Markov chain approximation when the utility function is exponential.

* Corresponding author.
*E-mail addresses:* wen.li@curtin.edu.au (W. Li), song.wang@uwa.edu.au (S. Wang).

http://dx.doi.org/10.1016/j.camwa.2017.03.024
0898-1221/© 2017 Elsevier Ltd. All rights reserved.

DISTRIBUTION A. Approved for public release: distribution unlimited.

In this paper, we will develop a new, efficient and accurate numerical method for computing European option prices based on the pricing model in [19,20]. In both [19,20], the utility function is assumed to be an exponential function. It is well known that using the exponential utility function can reduce the number of state variables in the HJB equation by one under a proper transformation. Thus, the use of this special function can simplify the problem considerably. Although the transformation substantially reduces the computational cost, it may not be applicable to other types of utility functions. The aim of this paper is to develop a numerical method which can efficiently and accurately solve the HJB equation without any dimension reduction technique. Therefore, the numerical method developed in this work can be used for computing option prices with any types of utility function.

This paper is organized as follows. In Section 2, we give a brief account of the formulation of the European option valuation problem as a set of HJB equations using the utility maximization theory. In Section 3, we first use a known penalty approach to approximate the HJB equations by a nonlinear PDE with penalty terms to penalize the parts which violate the constraints. We then propose a finite volume scheme for the penalty equation. In Section 4, an iterative algorithm and its convergence will be provided and in Section 5, we present the numerical results to demonstrate the usefulness of the numerical method.

## 2. The European option pricing model

In this section, we will present a brief account of the European option pricing model when the volatility is stochastic and trading the underling stocks is subject to proportional transaction costs. A detailed mathematical deduction of the model can be found in [16,20].

### 2.1. Stochastic volatility model with transaction costs

Consider a market consisting of a risky stock and a risk-less bond. Assume that the price of the stock at time $u \in [0, T]$, denoted as $S_u$, evolves according to the following stochastic volatility model:

$$\frac{dS_u}{S_u} = \mu du + \sqrt{v(u)}dW_u^1, \tag{1}$$

where $\mu$ is constant drift rate and $\sqrt{v(u)}$ is the volatility function which satisfies the following Cox–Ingersoll–Ross (CIR) process:

$$dv(u) = \xi(\eta - v(u))du + \vartheta\sqrt{v(u)}dW_u^2, \tag{2}$$

where $\xi$ is the speed of adjustment, $\eta$ is the mean and $\vartheta$ is the volatility to volatility. In (2) $\xi$, $\eta$ and $\vartheta$ are assumed to be constant satisfying $2\xi\eta > \vartheta^2$, and $W_u^1$ and $W_u^2$ are Wiener processes on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_u)_{0\le u\le T}, P)$ with correlation $\rho$.

We also assume that the price of the bond, $B(u)$, evolves according to the following ordinary differential equation

$$dB(u) = rB(u)du,$$

where $r \ge 0$ is a constant interest rate.

Suppose that the investors must pay transaction costs when buying or selling the stock and the transaction costs are proportional to the amount transferred from the stock to the bond. Let $\beta_u$ denote the amount the investors hold in the bond and $\alpha_u$ the number of shares of the stock held by the investors at time $u \in [0, T]$, then the evolution equations for $\beta_u$ and $\alpha_u$ are

$$d\beta_u = r\beta_u du - (1+\theta)S_u dL_u + (1-\theta)S_u dM_u, \tag{3}$$

$$d\alpha_u = dL_u - dM_u, \tag{4}$$

where $\theta \in [0, 1)$ represents the proportional transaction cost rate when buying and selling the stock, and $L_u$ and $M_u$ denote respectively the cumulative number of shares purchased and sold up to time $u$. Let $c(\alpha_u, S_u)$ denote the liquidated cash value of the stock and $W_u$ the investor's wealth at time $u$. We have

$$c(\alpha_u, S_u) = S_u(\alpha_u - \theta|\alpha_u|)$$
$$W_u(\alpha_u, \beta_u, S_u) = \beta_u + S_u(\alpha_u - \theta|\alpha_u|).$$

### 2.2. European option pricing via utility maximization

We now describe the utility based option pricing approach. The idea of the utility based option pricing approach is to consider an optimal portfolio selection problem of an investor whose objective is to find an admissible trading strategy to maximize his/her expect utility of terminal wealth. Under this approach, the reservation purchase (respectively write) price of an option is the price at which the investor has the same maximum expected utility whether he/she buys (respectively writes) the option or not. To use this approach to value reservation purchase and write prices of European call options, we first need to define the following three different utility maximization problems.

DISTRIBUTION A. Approved for public release: distribution unlimited.

**Problem 2.1** (*Utility Maximization for an Investor Without an Option*)**.** Consider an investor who trades only in the underlying stock and the bond. At time $t \in [0, T]$, the investor holds $\beta$ dollars in the bond and $\alpha$ shares of the stock of price $S$ with volatility $\nu$. The objective of the investor is to maximize the expected utility of terminal wealth over all admissible strategies, i.e.,

$$V^0(t, \alpha, \beta, S, \nu) = \sup_{\Lambda^0(t,\alpha,\beta,S,\nu)} E_t[U(W_T)] \quad (0 \le t \le T), \tag{5}$$

where $V^0(t, \alpha, \beta, S, \nu)$ denotes the investor's time $t$ maximum expected utility of terminal wealth (also known as value function), $E_t$ denotes the expectation operator conditional on the time $t$ information $(\alpha, \beta, S, \nu)$ and $U(.)$ is a utility function. $\Lambda^0(t, \alpha, \beta, S, \nu)$ is the set of admissible strategies available to the investor which is defined as the set of right-continuous, measurable, $F$-adapted, increasing processes, $L_u$ and $M_u$ $(t \le u \le T)$, such that the following conditions are satisfied:

1. The associated processes $(\alpha^{L_u,M_u}, \beta^{L_u,M_u}, S_u, \nu_u)$ satisfy (1)–(4) in $[t, T]$ with the initial state $(t, \alpha, \beta, S, \nu)$.
2. $\beta^{L_u,M_u} + S_u\alpha^{L_u,M_u} - S_u\theta|\alpha^{L_u,M_u}| > 0, \forall u \in [t, T]$.

The choice of the utility function $U$ is non-unique and a popular one is the following exponential function:

$$U(W) = 1 - \exp(-\gamma W), \tag{6}$$

where $\gamma > 0$ is a constant risk aversion parameter.

**Problem 2.2** (*Utility Maximization for an Investor Buying an Option*)**.** Assume that the investor trades in the market for the underlying stock and the bond, and in addition, purchases a cash-settled European call option written on the stock with strike price $K$ and expiry date $T$. Then the investor's objective is to choose an admissible trading strategy to maximize the expected utility of terminal wealth, i.e.,

$$V^b(t, \alpha, \beta, S, \nu) = \sup_{\Lambda^b(t,\alpha,\beta,S,\nu)} E_t[U(W_T + (S_T - K)^+)] \quad (0 \le t \le T), \tag{7}$$

where $\Lambda^b(t, \alpha, \beta, S, \nu) = \Lambda^0(t, \alpha, \beta, S, \nu)$ and $x^+ = \max\{x, 0\}$.

**Problem 2.3** (*Utility Maximization for an Investor Writing an Option*)**.** Assume that the investor trades in the market for the underlying stock and the bond, and, in addition, sells a cash-settled European call option written on the stock with strike price $K$ and expiry date $T$. Then, the investor's objective is to maximize the expected utility of terminal wealth over the set of feasible strategies, i.e.,

$$V^w(t, \alpha, \beta, S, \nu) = \sup_{\Lambda^w(t,\alpha,\beta,S,\nu)} E_t[U(W_T - (S_T - K)^+)] \quad (0 \le t \le T), \tag{8}$$

where $\Lambda^w(t, \alpha, \beta, S, \nu)$ denotes the writer's admissible strategies which are defined as the set of right-continuous, measurable, $F$-adapted, increasing processes, $L_u$ and $M_u$ $(t \le u \le T)$, such that the following conditions are satisfied.

1. The associated processes $(\alpha^{L_u,M_u}, \beta^{L_u,M_u}, S_u, \nu_u)$ satisfy (1)–(4) in $[t, T]$ with the initial state $(t, \alpha, \beta, S, \nu)$.
2. $\beta^{L_u,M_u} + S_u\left(\alpha^{L_u,M_u} - 1/(1-\theta)\right) - S_u\theta\left|\alpha^{L_u,M_u} - 1/(1-\theta)\right| > 0, \ \forall u \in [t, T]$.

We comment that Item 2 in each of Problems 2.1–2.3 represents the no-bankruptcy restriction. These conditions ensure that the investor's wealth is positive at all trading times.

Using the above problems, we now define the reservation purchase and write prices of a European call options as follows.

**Definition 2.4** (*Reservation Purchase Price of a European Call Option*)**.** Consider an investor who starts trading at time $t = 0$ with holding $\beta$ dollars in the bond and $\alpha$ shares of the stock of price $S$ with volatility $\nu$. Assume that the investor only can buy the option at the initial time $t = 0$. Then the investor's reservation purchase price of a European call option is defined as the amount, $P_b$, such that $V^b(0, \alpha, \beta - P_b, S, \nu) = V^0(0, \alpha, \beta, S, \nu)$.

**Definition 2.5** (*Reservation Write Price of a European Call Option*)**.** Consider an investor who starts trading at time $t = 0$ with holding $\beta$ dollars in the bond and $\alpha$ shares of the stock whose price is $S$ with initial volatility $\nu$. Assume that the investor can only sell the option at the initial time $t = 0$. Then the investor's reservation write price of a European call option is defined as the amount, $P_w$, such that $V^w(0, \alpha, \beta + P_w, S, \nu) = V^0(0, \alpha, \beta, S, \nu)$.

From the above definitions it is clear that computing reservation purchase or write price of a European option involves two of the three value functions defined in (5)–(8). By the dynamic programming principle, we can derive an HJB equation with a set of appropriate terminal conditions governing the value functions $V^0$, $V^b$ and $V^w$.

Let $\mathcal{L}_k$, $k = 1, 2, 3$, be the linear differential operators defined respectively by

$$\mathcal{L}_1 = -\left(\frac{\partial}{\partial t} + r\beta\frac{\partial}{\partial \beta} + \mu S\frac{\partial}{\partial S} + \xi(\eta - \nu)\frac{\partial}{\partial \nu} + \frac{1}{2}S^2\nu\frac{\partial^2}{\partial S^2} + \frac{1}{2}\vartheta^2\nu\frac{\partial^2}{\partial \nu^2} + \rho\vartheta S\nu\frac{\partial^2}{\partial S\partial \nu}\right), \tag{9}$$

DISTRIBUTION A. Approved for public release: distribution unlimited.

$$\mathcal{L}_2 = -\frac{\partial}{\partial \alpha} + (1+\theta)S\frac{\partial}{\partial \beta}, \tag{10}$$

$$\mathcal{L}_3 = \frac{\partial}{\partial \alpha} - (1-\theta)S\frac{\partial}{\partial \beta}. \tag{11}$$

Then, the value functions $V^0, V^b, V^w$ are defined by the following HJB equation

$$\min\{\mathcal{L}_1 V, \mathcal{L}_2 V, \mathcal{L}_3 V\} = 0, \tag{12}$$

for $(t, \alpha, \beta, S, \nu) \in [0, T) \times \Omega^i \times (0, +\infty)$ satisfying, respectively, the following terminal conditions:

$$V(T, \alpha, \beta, S, \nu) = V^i(T, \alpha, \beta, S, \nu), \quad (\alpha, \beta, S, \nu) \in \Omega^i \times (0, +\infty) \tag{13}$$

for $i = 0, b$ and $w$ respectively, where

$$V^0(T, \alpha, \beta, S, \nu) = U(\beta + S(\alpha - \theta|\alpha|)), \tag{14}$$

$$V^b(T, \alpha, \beta, S, \nu) = U(\beta + S(\alpha - \theta|\alpha|) + (S-K)^+), \tag{15}$$

$$V^w(T, \alpha, \beta, S, \nu) = U(\beta + S(\alpha - \theta|\alpha|) - (S-K)^+), \tag{16}$$

and

$$\Omega^0 = \Omega^b = \{(\alpha, \beta, S) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+ : \beta + S\alpha - S\theta|\alpha| > 0\}, \tag{17}$$

$$\Omega^w = \left\{(\alpha, \beta, S) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+ : \beta + S\left(\alpha - 1/(1-\theta)\right) - S\theta\left|\alpha - 1/(1-\theta)\right| > 0\right\}. \tag{18}$$

In the above $(v)^+ = \max\{v, 0\}$. Note that (12) is nonlinear and it does not have in general classical solutions. It has been proved in Cosso et al. [20] that the value functions defined by (5)–(8) are unique viscosity solutions of their respective HJB equations (12)–(16). This is given in the following theorem.

**Theorem 2.6.** *Let $i \in \{0, b, w\}$ and assume that the value function $V^i$ is continuous on $[0, T] \times \bar{\Omega}^i \times (0, +\infty)$, then the value function $V^i$ is the unique constrained viscosity solution of (12) with the terminal condition*

$$V(T, \alpha, \beta, S, \nu) = V^i(T, \alpha, \beta, S, \nu), \quad \text{for } (\alpha, \beta, S, \nu) \in \Omega^i \times (0, +\infty),$$

*where $V^i(T, \alpha, \beta, S, \nu)$ and $\Omega^i$ are defined in (14)–(16) and (17)–(18) respectively for $i = 0, b, w$.*

## 3. The numerical techniques

Note that (12) can be regarded as a constrained optimization problem in infinite dimensions. Thus, the numerical solution of (12) involves numerical optimization techniques and discretization schemes. In this section, we will propose a nonlinear PDE containing penalty terms, called penalty equation, to approximate (12). The penalty terms in the penalty equation penalize the part of its solution which violate $\mathcal{L}_i V \geq 0$ for $i = 2, 3$, while $\mathcal{L}_1 V \geq 0$ is automatically satisfied by the formulation. We will then develop a finite volume method along with the full implicit 2-step time stepping method for the resulting penalty equation and show that the system matrix is an $M$-matrix.

### 3.1. The penalty approach

Penalty methods have been developed for solving both finite- and infinite-dimensional HJB equations [21–25]. In particular, we propose a penalty method in [16–18] for the European and American option pricing problems under proportional transaction costs with a constant volatility. Motivated by our previous work, we propose the following penalty formulation for (12):

$$\mathcal{L}_1 V_\lambda + \lambda[\mathcal{L}_2 V_\lambda]^- + \lambda[\mathcal{L}_3 V_\lambda]^- = 0 \tag{19}$$

for $(t, \alpha, \beta, S, \nu) \in [0, T) \times \Omega^i \times (0, +\infty)$ with the terminal condition

$$V_\lambda(T, \alpha, \beta, S, \nu) = V^i(T, \alpha, \beta, S, \nu), \quad \text{for } (\alpha, \beta, S, \nu) \in \Omega^i \times (0, +\infty), \tag{20}$$

where $\mathcal{L}_k$ are the differential operators defined in (9)–(11), $V^i(T, \alpha, \beta, S, \nu)$ is the boundary condition given in (14)–(16) for each $i$, $\lambda > 1$ is a penalty parameter, $i \in \{0, b, w\}$ and $(v)^- = \min\{v, 0\}$ for any function $v$.

For the solution of (19) we have the following convergence result.

**Theorem 3.1.** *For any $i \in \{0, b, w\}$, let $V^i$ be the unique constrained viscosity solution of (12)–(13). For each $\lambda > 1$, (19)–(20) has a unique viscosity solution $v^i_\lambda$ and $v^i_\lambda \to V^i$ as $\lambda \to \infty$.*

The proof of this theorem is essentially a repetition of the proofs of Theorems 4.2 and 4.3 in [16] in which $\mathcal{L}_1$ is a special case of that defined in (9). However, all the required properties used in the proofs in [16] are satisfied by $\mathcal{L}_1$ in (9). Thus, we omit this proof.

DISTRIBUTION A. Approved for public release: distribution unlimited.

### 3.2. The finite volume method

Finite volume methods have been used for solving one and two-dimensional Black–Scholes equations [26–28]. In this section, we will propose a finite volume method with an upwind technique for (19). This method has the merit that the coefficient matrix of the resulting system matrix from the method is always an $M$-matrix even when $\rho \neq 0$ in (9). In fact, this property cannot be achieved by any finite difference discretization scheme. For brevity, we only consider the case that $i = b$ in (19) and (20). The methods for the other two cases are essentially the same as that for $i = b$ and thus are omitted. Before proceeding, we first rewrite the second and last terms of (19) as the following equivalent form:

$$\lambda[\mathcal{L}_2 V_\lambda]^- = \min_{\bar{m}\in[0,\lambda]} \bar{m}\mathcal{L}_2 V_\lambda, \qquad \lambda[\mathcal{L}_3 V_\lambda]^- = \min_{\bar{n}\in[0,\lambda]} \bar{n}\mathcal{L}_3 V_\lambda.$$

Then, (19) can be rewritten as the divergence form as:

$$-\frac{\partial V}{\partial t} - \nabla\cdot(A\nabla V) + \underline{b}\cdot\nabla V + \min_{\bar{m}\in[0,\lambda]} \bar{m}\mathcal{L}_2 V_\lambda + \min_{\bar{n}\in[0,\lambda]} \bar{n}\mathcal{L}_3 V_\lambda = 0, \tag{21}$$

where

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & a_{33} & a_{34} \\ 0 & 0 & a_{43} & a_{44} \end{pmatrix} := \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2}S^2\nu & \frac{1}{2}\rho\vartheta S\nu \\ 0 & 0 & \frac{1}{2}\rho\vartheta S\nu & \frac{1}{2}\vartheta^2\nu \end{pmatrix}, \tag{22}$$

$$\underline{b} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix} := \begin{pmatrix} 0 \\ -r\beta \\ S\nu + \frac{1}{2}\rho\vartheta S - \mu S \\ \frac{1}{2}\rho\vartheta\nu + \frac{1}{2}\vartheta^2 - \xi(\eta - \nu) \end{pmatrix}. \tag{23}$$

Let $\Omega = (-\infty, +\infty)^2 \times (0, +\infty)^2$. We consider the problem in the following finite region:

$$\Omega_L := (-\underline{L}_\alpha, \bar{L}_\alpha) \times (-\underline{L}_\beta, \bar{L}_\beta) \times (0, \bar{L}_S) \times (0, \bar{L}_\nu) \subset \Omega \tag{24}$$

where $\underline{L}_\alpha$, $\underline{L}_\beta$, $\bar{L}_\alpha$, $\bar{L}_\beta$, $\bar{L}_S$ and $\bar{L}_\nu$ are positive constants. To discretize $\Omega_L$, we choose four positive integers $E$, $M$, $P$ and $Z$ and use these integers to define a uniform mesh for $\Omega_L$ with mesh nodes

$$\begin{aligned} &\alpha_i = -\underline{L}_\alpha + i \times h_1, \quad i = 0, 1, 2, \ldots, E, \\ &\beta_j = -\underline{L}_\beta + j \times h_2, \quad j = 0, 1, 2, \ldots, M, \\ &S_k = k \times h_3, \quad k = 0, 1, 2, \ldots, P, \\ &\nu_l = l \times h_4, \quad l = 0, 1, 2, \ldots, Z, \end{aligned}$$

where

$$h_1 = \frac{\bar{L}_\alpha + \underline{L}_\alpha}{E}, \qquad h_2 = \frac{\bar{L}_\beta + \underline{L}_\beta}{M}, \qquad h_3 = \frac{\bar{L}_S}{P}, \qquad h_4 = \frac{\bar{L}_\nu}{Z}.$$

Let $h = \max\{h_1, h_2, h_3, h_4\}$. In what follows, we will characterize this spatial mesh using the grid index set: $\bar{G}_h = \partial G_h \cup G_h$, where $G_h$ and $\partial G_h$ denote the index sets of the interior and boundary mesh nodes defined respectively by

$$G_h = \{(i, j, k, l) : \ i = 1, 2, \ldots, E-1, j = 1, 2, \ldots, M-1, k = 1, 2, \ldots, P-1, l = 1, 2, \ldots, Z-1\}$$

and

$$\begin{aligned} \partial G_h = \{&(0, j, k, l), (E, j, k, l), (i, 0, k, l), (i, M, k, l), (i, j, 0, l), (i, j, P, l), \\ &(i, j, k, 0), (i, j, k, l) : \ i = 1, 2 \ldots E, j = 1, 2, \ldots, M, k = 1, 2, \ldots, P, l = 1, 2, \ldots, Z\}. \end{aligned}$$

Clearly, each grid point $(i, j, k, l) \in \bar{G}_h$ corresponds to a state $(\alpha_i, \beta_j, S_k, \nu_l)$.

Dual to the above mesh (called primary mesh), we define a secondary mesh with the mesh nodes

$$\alpha_{i-\frac{1}{2}} = \frac{\alpha_{i-1} + \alpha_i}{2}, \qquad \beta_{j-\frac{1}{2}} = \frac{\beta_{j-1} + \beta_j}{2}, \qquad S_{k-\frac{1}{2}} = \frac{S_{k-1} + S_k}{2}, \qquad \nu_{l-\frac{1}{2}} = \frac{\nu_{l-1} + \nu_l}{2},$$

DISTRIBUTION A. Approved for public release: distribution unlimited.

for $i = 0, \ldots, E+1, j = 0, \ldots, M+1, k = 0, \ldots, P+1$ and $l = 0, \ldots, Z+1$ with the convention

$$\alpha_{-\frac{1}{2}} = \alpha_0, \quad \alpha_{E+\frac{1}{2}} = \alpha_E, \quad \beta_{-\frac{1}{2}} = \beta_0, \quad \alpha_{M+\frac{1}{2}} = \beta_M,$$
$$S_{-\frac{1}{2}} = S_0, \quad S_{P+\frac{1}{2}} = S_P, \quad \nu_{-\frac{1}{2}} = \nu_0, \quad \nu_{Z+\frac{1}{2}} = \nu_Z.$$

For each mesh node $(i, j, k, l) \in \bar{G}_h$, we define a so-called box or control region centered at the point by

$$R_{ijkl} = \left(\alpha_{i-\frac{1}{2}}, \alpha_{i+\frac{1}{2}}\right) \times \left(\beta_{j-\frac{1}{2}}, \beta_{j+\frac{1}{2}}\right) \times \left(S_{k-\frac{1}{2}}, S_{k+\frac{1}{2}}\right) \times \left(\nu_{l-\frac{1}{2}}, \nu_{l+\frac{1}{2}}\right).$$

Integrating (21) over each control region $R_{ijkl}$ and applying integration by parts to the 2nd term, we have

$$-\int_{R_{ijkl}} \frac{\partial V}{\partial t} d\alpha d\beta dS d\nu - \int_{\partial R_{ijkl}} (A\nabla V) \cdot \underline{n} d\sigma + \int_{R_{ijkl}} \underline{b} \cdot (\nabla V) d\alpha d\beta dS d\nu$$
$$+ \int_{R_{ijkl}} \min_{\bar{m}\in[0,\lambda]} \bar{m}\mathcal{L}_2 V_\lambda d\alpha d\beta dS d\nu + \int_{R_{ijkl}} \min_{\bar{n}\in[0,\lambda]} \bar{n}\mathcal{L}_3 V_\lambda d\alpha d\beta dS d\nu = 0 \tag{25}$$

for $(i, j, k, l) \in G_h$, where $\partial R_{ijkl}$ denotes the boundary of $R_{ijkl}$, $\underline{n}$ the unit vector out-normal to $\partial R_{ijkl}$ and $d\sigma$ denotes the 3d infinitesimal along $\partial R_{ijkl}$. Using the 1-point quadrature rule and (23), we have

$$\int_{R_{ijkl}} \frac{\partial V}{\partial t} d\alpha d\beta dS\, d\nu \approx \frac{\partial V_{ijkl}}{\partial t} |R_{ijkl}|, \tag{26}$$

$$\int_{R_{ijkl}} \underline{b} \cdot (\nabla V) d\alpha d\beta dS d\nu \approx \left(b_2 \frac{\partial V}{\partial \beta} + b_3 \frac{\partial V}{\partial S} + b_4 \frac{\partial V}{\partial \nu}\right)_{(\alpha_i, \beta_j, S_k, \nu_l)} |R_{ijkl}|, \tag{27}$$

$$\int_{R_{ijkl}} \min_{\bar{m}\in[0,\lambda]} \bar{m}\mathcal{L}_2 V_\lambda d\alpha d\beta dS d\nu \approx \min_{\bar{m}\in[0,\lambda]} \bar{m} \left(-\frac{\partial V}{\partial \alpha} + (1+\theta) S \frac{\partial V}{\partial \beta}\right)_{(\alpha_i, \beta_j, S_k, \nu_l)} |R_{ijkl}|, \tag{28}$$

$$\int_{R_{ijkl}} \min_{\bar{n}\in[0,\lambda]} \bar{n}\mathcal{L}_3 V_\lambda d\alpha d\beta dS d\nu \approx \min_{\bar{n}\in[0,\lambda]} \bar{n} \left(\frac{\partial V}{\partial \alpha} - (1-\theta) S \frac{\partial V}{\partial \beta}\right)_{(\alpha_i, \beta_j, S_k, \nu_l)} |R_{ijkl}|, \tag{29}$$

where $|\cdot|$ denotes the 'measure' (absolute value, area or volume depending on the context) of a quantity and $V_{ijkl}$ denote an approximation of $V$ at the mesh node.

We now consider the approximation of the second term in (25). Since $R_{ijk}$ is a hyper-rectangle or box in 4D, $\partial R_{ijkl}$ contains 8 3D rectangular prisms or facets. Each of these facets is perpendicular to one of the axes so that its normal direction $\underline{n}$ is in or opposite the direction of the axis. In fact, the possible normal directions are $(\pm 1, 0, 0, 0)^\top$, $(0, \pm 1, 0, 0)^\top$, $(0, 0, \pm 1, 0)^\top$ and $(0, 0, 0, \pm 1)^\top$. From the definition $A$ in (22) and these choices of $\underline{n}$ we see that $A\nabla V \cdot \underline{n}$ has only 4 non-zero terms corresponding to the facets intersecting $S_{k\pm\frac{1}{2}}$ and $\nu_{l\pm\frac{1}{2}}$. Therefore, we have

$$-\int_{\partial R_{ijkl}} A\nabla V \cdot \underline{n} d\sigma = -\int_{\partial R^1_{ijkl}} \left(a_{33}\frac{\partial V}{\partial S} + a_{34}\frac{\partial V}{\partial \nu}\right) d\alpha d\beta d\nu + \int_{\partial R^3_{ijkl}} \left(a_{33}\frac{\partial V}{\partial S} + a_{34}\frac{\partial V}{\partial \nu}\right) d\alpha d\beta d\nu$$
$$- \int_{\partial R^2_{ijkl}} \left(a_{43}\frac{\partial V}{\partial S} + a_{44}\frac{\partial V}{\partial \nu}\right) d\alpha d\beta dS + \int_{\partial R^4_{ijkl}} \left(a_{43}\frac{\partial V}{\partial S} + a_{44}\frac{\partial V}{\partial \nu}\right) d\alpha\, d\beta dS,$$

where $\partial R^m_{ijkl}$, $m = 1, 2, 3, 4$, denote the 4 facets of $\partial R_{ijkl}$ on which $A\nabla V \cdot \underline{n} \neq 0$. Applying the 1-point quadrature rule to the above equation and noting that these facets are numbered in such a way that $|R^1_{ijkl}| = |R^3_{ijkl}|$ and $|R^2_{ijkl}| = |R^4_{ijkl}|$, we have

$$\begin{aligned} -\int_{\partial R_{ijkl}} A\nabla V \cdot \underline{n} d\sigma \approx & -\left(a_{33}\frac{\partial V}{\partial S} + a_{34}\frac{\partial V}{\partial \nu}\right)\Big|_{\left(\alpha_i, \beta_j, S_{k+\frac{1}{2}}, \nu_l\right)} \times |R^3_{ijkl}| \\ & + \left(a_{33}\frac{\partial V}{\partial S} + a_{34}\frac{\partial V}{\partial \nu}\right)\Big|_{\left(\alpha_i, \beta_j, S_{k-\frac{1}{2}}, \nu_l\right)} \times |R^3_{ijkl}| \\ & - \left(a_{43}\frac{\partial V}{\partial S} + a_{44}\frac{\partial V}{\partial \nu}\right)\Big|_{\left(\alpha_i, \beta_j, S_k, \nu_{l+\frac{1}{2}}\right)} \times |R^4_{ijkl}| \\ & + \left(a_{43}\frac{\partial V}{\partial S} + a_{44}\frac{\partial V}{\partial \nu}\right)\Big|_{\left(\alpha_i, \beta_j, S_k, \nu_{l-\frac{1}{2}}\right)} \times |R^4_{ijkl}|, \end{aligned} \tag{30}$$

DISTRIBUTION A. Approved for public release: distribution unlimited.

where

$$R^3_{ijkl} = \left(\alpha_{i-\frac{1}{2}}, \alpha_{i+\frac{1}{2}}\right) \times \left(\beta_{j-\frac{1}{2}}, \beta_{j+\frac{1}{2}}\right) \times \left(\nu_{l-\frac{1}{2}}, \nu_{l+\frac{1}{2}}\right),$$
$$R^4_{ijkl} = \left(\alpha_{i-\frac{1}{2}}, \alpha_{i+\frac{1}{2}}\right) \times \left(\beta_{j-\frac{1}{2}}, \beta_{j+\frac{1}{2}}\right) \times \left(S_{k-\frac{1}{2}}, S_{k+\frac{1}{2}}\right).$$

Replacing the terms in (21) by their respective approximations in (26)–(30), we have

$$\begin{aligned}
&-\frac{\partial V_{ijkl}}{\partial t}|R_{ijkl}| - \left(a_{33}\frac{\partial V}{\partial S} + a_{34}\frac{\partial V}{\partial \nu}\right)_{\left(\alpha_i,\beta_j,S_{k+\frac{1}{2}},\nu_l\right)}|R^3_{ijkl}| + \left(a_{33}\frac{\partial V}{\partial S} + a_{34}\frac{\partial V}{\partial \nu}\right)_{\left(\alpha_i,\beta_j,S_{k-\frac{1}{2}},\nu_l\right)}|R^3_{ijkl}| \\
&- \left(a_{43}\frac{\partial V}{\partial S} + a_{44}\frac{\partial V}{\partial \nu}\right)_{\left(\alpha_i,\beta_j,S_k,\nu_{l+\frac{1}{2}}\right)}|R^4_{ijkl}| + \left(a_{43}\frac{\partial V}{\partial S} + a_{44}\frac{\partial V}{\partial \nu}\right)_{\left(\alpha_i,\beta_j,S_k,\nu_{l-\frac{1}{2}}\right)}|R^4_{ijkl}| \\
&+ \left(b_2\frac{\partial V}{\partial \beta} + b_3\frac{\partial V}{\partial S} + b_4\frac{\partial V}{\partial \nu}\right)_{(\alpha_i,\beta_j,S_k,\nu_l)}|R_{ijkl}| + \min_{\bar{m}\in[0,\lambda]}\bar{m}\left(-\frac{\partial V}{\partial \alpha} + (1+\theta)S\frac{\partial V}{\partial \beta}\right)\Bigg|_{(\alpha_i,\beta_j,S_k,\nu_l)}|R_{ijkl}| \\
&+ \min_{\bar{n}\in[0,\lambda]}\bar{n}\left(\frac{\partial V}{\partial \alpha} - (1-\theta)S\frac{\partial V}{\partial \beta}\right)\Bigg|_{(\alpha_i,\beta_j,S_k,\nu_l)}|R_{ijkl}| = 0.
\end{aligned} \tag{31}$$

Given a positive integer $N$, we divide the time interval $[0, T]$ into $N$ sub-intervals with time points $t_n = n \times \Delta t$ for $n = 0, 1, \ldots, N$, where $\Delta t = T/N$. We let $G_{\Delta t} = \{0, 1, 2, \ldots, N\}$ denote the index set of the time mesh points.

We now approximate the 1st spatial derivatives in (31) by finite-differences. For any admissible $(n, i, j, k, l)$, we denote by $V^n_{ijkl}$ the approximation (to be determined) the solution to (31) and (20) at the node $(t_n, \alpha_i, \beta_j, S_k, \nu_l)$. Using the following finite difference operators

$$\begin{aligned}
&D_t V^n_{ijkl} = \frac{V^{n+1}_{ijkl} - V^n_{ijkl}}{\Delta t}, \\
&D^+_\alpha V^n_{ijkl} = \frac{V^n_{(i+1)jkl} - V^n_{ijkl}}{h_1}, \qquad D^-_\alpha V^n_{ijkl} = \frac{V^n_{ijkl} - V^n_{(i-1)jkl}}{h_1}, \\
&D^+_\beta V^n_{ijkl} = \frac{V^n_{i(j+1)kl} - V^n_{ijkl}}{h_2}, \qquad D^-_\beta V^n_{ijkl} = \frac{V^n_{ijkl} - V^n_{i(j-1)kl}}{h_2}, \\
&D^+_S V^n_{ijkl} = \frac{V^n_{ij(k+1)l} - V^n_{ijkl}}{h_3}, \qquad D^-_S V^n_{ijkl} = \frac{V^n_{ijkl} - V^n_{ij(k-1)l}}{h_4} \\
&D^+_\nu V^n_{ijkl} = \frac{V^n_{ijk(l+1)} - V^n_{ijkl}}{h_4}, \qquad D^-_\nu V^n_{ijkl} = \frac{V^n_{ijkl} - V^n_{ijk(l-1)}}{h_4}, \\
&D_S V^n_{ij\left(k+\frac{1}{2}\right)l} = \frac{V^n_{ij(k+1)l} - V^n_{ijkl}}{h_3}, \qquad D_S V^n_{ij\left(k-\frac{1}{2}\right)l} = \frac{V^n_{ijkl} - V^n_{ij(k-1)l}}{h_3}, \\
&D_\nu V^n_{ijk\left(l+\frac{1}{2}\right)} = \frac{V^n_{ijk(l+1)} - V^n_{ijkl}}{h_4}, \qquad D_\nu V^n_{ijk\left(l-\frac{1}{2}\right)} = \frac{V^n_{ijkl} - V^n_{ijk(l-1)}}{h_3},
\end{aligned}$$

we propose a finite difference scheme for (31) as follows:

$$\begin{aligned}
&-D_t V^n_{ijkl}|R_{ijkl}| - \frac{1}{2}S^2_{k+\frac{1}{2}}\nu_l D_S V^n_{ij\left(k+\frac{1}{2}\right)l}|R^3_{ijkl}| - \frac{1}{2}\left(\rho\vartheta S_{k+\frac{1}{2}}\nu_l\right)^+ D^+_\nu V^n_{ijkl}|R^3_{ijkl}| \\
&- \frac{1}{2}\left(\rho\vartheta S_{k+\frac{1}{2}}\nu_l\right)^- D^-_\nu V^n_{ijkl}|R^3_{ijkl}| + \left(\frac{1}{2}S^2_{k-\frac{1}{2}}\nu_l\right) D_S V^n_{ij\left(k-\frac{1}{2}\right)l}|R^3_{ijkl}| \\
&+ \frac{1}{2}\left(\rho\vartheta S_{k-\frac{1}{2}}\nu_l\right)^+ D^-_\nu V^n_{ijkl}|R^3_{ijkl}| + \frac{1}{2}\left(\rho\vartheta S_{k-\frac{1}{2}}\nu_l\right)^- D^+_\nu V^n_{ijkl}|R^3_{ijkl}| \\
&- \frac{1}{2}\left(\rho\vartheta S_k\nu_{l+\frac{1}{2}}\right)^+ D^+_S V^n_{ijkl}|R^4_{ijkl}| - \frac{1}{2}\left(\rho\vartheta S_k\nu_{l+\frac{1}{2}}\right)^- D^-_S V^n_{ijkl}|R^4_{ijkl}| \\
&- \left(\frac{1}{2}\vartheta^2_k\nu_{l+\frac{1}{2}}\right) D_\nu V^n_{ijk\left(l+\frac{1}{2}\right)}|R^4_{ijkl}| + \frac{1}{2}\left(\rho\vartheta S_k\nu_{l-\frac{1}{2}}\right)^+ D^-_S V^n_{ijkl}|R^4_{ijkl}| \\
&+ \frac{1}{2}\left(\rho\vartheta S_k\nu_{l-\frac{1}{2}}\right)^- D^+_S V^n_{ijkl}|R^4_{ijkl}| + \left(\frac{1}{2}\vartheta^2_k\nu_{l-\frac{1}{2}}\right) D_\nu V^n_{ijk\left(l-\frac{1}{2}\right)}|R^4_{ijkl}|
\end{aligned}$$

DISTRIBUTION A. Approved for public release: distribution unlimited.

$$
\begin{aligned}
&- r(\beta_j)^+ D_\beta^+ V_{ijkl}^n |R_{ijkl}| - r(\beta_j)^- D_\beta^- V_{ijkl}^n |R_{ijkl}| + \left(S_k \nu_l + \frac{1}{2}\rho\vartheta S_k - \mu S_k\right)^+ D_S^- V_{ijkl}^n |R_{ijkl}| \\
&+ \left(S_k \nu_l + \frac{1}{2}\rho\vartheta S_k - \mu S_k\right)^- D_S^+ V_{ijkl}^n |R_{ijkl}| + \left(\frac{1}{2}\rho\vartheta \nu_l + \frac{1}{2}\vartheta^2 + \xi(\eta - \nu_l)\right)^+ D_\nu^- V_{ijkl}^n |R_{ijkl}| \\
&+ \left(\frac{1}{2}\rho\vartheta \nu_l + \frac{1}{2}\vartheta^2 + \xi(\eta - \nu_l)\right)^- D_\nu^+ V_{ijkl}^n |R_{ijkl}| + m_1(n,i,j,k,l)\left(-D_\alpha^+ V_{ijkl}^n + (1+\theta)S_k D_\beta^- V_{ijkl}^n\right)|R_{ijkl}| \\
&+ n_1(n,i,j,k,l)\left(D_\alpha^- V_{ijkl}^n - (1-\theta)S_k D_\beta^+ V_{ijkl}^n\right)|R_{ijkl}| = 0,
\end{aligned} \tag{32}
$$

where $(\cdot)^+ = \max\{\cdot, 0\}$ and $(\cdot)^- = \min\{\cdot, 0\}$ are as defined before and

$$
m_1(n,i,j,k,l) = \arg\min_{\bar{m}\in[0,\lambda]} \bar{m}\left(-\frac{V_{(i+1)jkl}^n - V_{ijkl}^n}{h_1} + (1+\theta)S_k \frac{V_{ijkl}^n - V_{i(j-1)kl}^n}{h_2}\right), \tag{33}
$$

$$
n_1(n,i,j,k,l) = \arg\min_{\bar{n}\in[0,\lambda]} \bar{n}\left(\frac{V_{ijkl}^n - V_{(i-1)jkl}^n}{h_1} - (1-\theta)S_k \frac{V_{i(j+1)kl}^n - V_{ijkl}^n}{h_2}\right). \tag{34}
$$

Note that in the original problem, only final/pay-off conditions are fined defined in (14)–(16). However, in computation, we need to define some artificial boundary conditions. For a detailed discussion on artificial boundary conditions, we refer to [29]. Using (20), (17) and (18), we define the boundary and terminal conditions for (32)–(34) as follows.

$$
V_{ijkl}^N = \begin{cases} U(\beta_j + S_k(\alpha_i - \theta|\alpha_i|) + (S_k - K)^+), & (i,j,k,l) \in G_h, x_{i,j,k} \in \bar{\Omega}^b, \\ 0, & (i,j,k,l) \in G_h, x_{i,j,k} \notin \bar{\Omega}^b, \end{cases} \tag{35}
$$

$$
V_{ijk}^n = \begin{cases} U(\beta_j + S_k(\alpha_i - \theta|\alpha_i|) + (S_k - K)^+), & (i,j,k,l) \in \partial G_h, x_{i,j,k} \in \bar{\Omega}^b, \\ 0, & (i,j,k,l) \in \partial G_h, x_{i,j,k} \notin \bar{\Omega}^b, \end{cases} \tag{36}
$$

for $n = 0, 1, \ldots, N$ where $x_{i,j,k} = (\alpha_i, \beta_j, S_k)$ and $U(W)$ is the utility function.

Let $\Delta = (\Delta t, h)$. For the discretization scheme defined in (32) we have the following theorem.

**Theorem 3.2.** *For any $\Delta > (0,0)$ and given $m_1(n,i,j,k,l) \in [0,\lambda]$ and $n_1(n,i,j,k,l) \in [0,\lambda]$, the system matrix of (32) is an M-matrix and the solution of (32) is bounded uniformly on $G_h \times G_{\Delta t}$.*

**Proof.** We first rewrite (32) in the following equivalent form:

$$
\begin{aligned}
&- \frac{V_{ijkl}^{n+1} - V_{ijkl}^n}{\Delta t} h_1h_2h_3h_4 - \frac{1}{2}S_{k+\frac{1}{2}}^2 \nu_l \frac{V_{ij(k+1)l}^n - V_{ijkl}^n}{h_3} h_1h_2h_4 \\
&- \frac{1}{2}\left(\rho\vartheta S_{k+\frac{1}{2}}\nu_l\right)^+ \frac{V_{ijk(l+1)}^n - V_{ijkl}^n}{h_4} h_1h_2h_4 + \frac{1}{2}\left|\left(\rho\vartheta S_{k+\frac{1}{2}}\nu_l\right)^-\right| \frac{V_{ijkl}^n - V_{ijk(l-1)}^n}{h_4} h_1h_2h_4 \\
&+ \frac{1}{2}S_{k-\frac{1}{2}}^2 \nu_l \frac{V_{ijkl}^n - V_{ij(k-1)l}^n}{h_3} h_1h_2h_4 + \frac{1}{2}\left(\rho\vartheta S_{k-\frac{1}{2}}\nu_l\right)^+ \frac{V_{ijkl}^n - V_{ijk(l-1)}^n}{h_4} h_1h_2h_4 \\
&- \frac{1}{2}\left|\left(\rho\vartheta S_{k-\frac{1}{2}}\nu_l\right)^-\right| \frac{V_{ijk(l+1)}^n - V_{ijkl}^n}{h_4} \times h_1h_2h_4 - \frac{1}{2}\left(\rho\vartheta S_k \nu_{l+\frac{1}{2}}\right)^+ \frac{V_{ij(k+1)l}^n - V_{ijkl}^n}{h_3} h_1h_2h_3 \\
&+ \frac{1}{2}\left|\left(\rho\vartheta S_k \nu_{l+\frac{1}{2}}\right)^-\right| \frac{V_{ijkl}^n - V_{ij(k-1)l}^n}{h_3} h_1h_2h_3 - \frac{1}{2}\vartheta_k^2 \nu_{l+\frac{1}{2}} \frac{V_{ijk(l+1)}^n - V_{ijkl}^n}{h_4} h_1h_2h_3 \\
&+ \frac{1}{2}\left(\rho\vartheta S_k \nu_{l-\frac{1}{2}}\right)^+ \frac{V_{ijkl}^n - V_{ij(k-1)l}^n}{h_3} \times h_1h_2h_3 - \frac{1}{2}\left|\left(\rho\vartheta S_k \nu_{l-\frac{1}{2}}\right)^-\right| \frac{V_{ij(k+1)l}^n - V_{ijkl}^n}{h_3} h_1h_2h_3 \\
&+ \frac{1}{2}\vartheta_k^2 \nu_{l-\frac{1}{2}} \frac{V_{ijkl}^n - V_{ijk(l-1)}^n}{h_4} h_1h_2h_3 - r(\beta_j)^+ \frac{V_{i(j+1)kl}^n - V_{ijkl}^n}{h_2} h_1h_2h_3h_4 \\
&+ r|(\beta_j)^-| \frac{V_{ijkl}^n - V_{i(j-1)kl}^n}{h_2} h_1h_2h_3h_4 + \left(S_k\nu_l + \frac{1}{2}\rho\vartheta S_k - \mu S_k\right)^+ \frac{V_{ijkl}^n - V_{ij(k-1)l}^n}{h_3} h_1h_2h_3h_4 \\
&- \left|\left(S_k\nu_l + \frac{1}{2}\rho\vartheta S_k - \mu S_k\right)^-\right| \frac{V_{ij(k+1)l}^n - V_{ijkl}^n}{h_3} h_1h_2h_3h_4 \\
&+ \left(\frac{1}{2}\rho\vartheta\nu_l + \frac{1}{2}\vartheta^2 - \xi(\eta - \nu_l)\right)^+ \frac{V_{ijkl}^n - V_{ijk(l-1)}^n}{h_4} h_1h_2h_3h_4
\end{aligned}
$$

DISTRIBUTION A. Approved for public release: distribution unlimited.

$$
\begin{aligned}
&- \left|\left(\frac{1}{2}\rho\vartheta\nu_l + \frac{1}{2}\vartheta^2 - \xi(\eta-\nu_l)\right)\right|^- \left|\frac{V^n_{ijk(l+1)} - V^n_{ijkl}}{h_4}\right| h_1h_2h_3h_4 \\
&+ m_1(n,i,j,k,l)\Big(-\frac{V^n_{(i+1)jkl} - V^n_{ijkl}}{h_1} + (1+\theta)S_k\frac{V^n_{ijkl} - V^n_{i(j-1)kl}}{h_2}\Big)h_1h_2h_3h_4 \\
&+ n_1(n,i,j,k,l)\Big(\frac{V^n_{ijkl} - V^n_{(i-1)jkl}}{h_1} - (1-\theta)S_k\frac{V^n_{i(j+1)kl} - V^n_{ijkl}}{h_2}\Big)h_1h_2h_3h_4 = 0.
\end{aligned} \tag{37}
$$

Multiplying (37) by $\frac{\Delta t}{h_1h_2h_3h_4}$ and rearranging the resulting equation, we have the following system:

$$
\begin{aligned}
V^{n+1}_{ijkl} = {} & V^n_{ijkl}\left[1 + \frac{1}{2}S^2_{k+\frac{1}{2}}\nu_l\frac{\Delta t}{h_3^2} + \frac{1}{2}\left|\left(\rho\vartheta S_{k+\frac{1}{2}}\nu_l\right)\right|\frac{\Delta t}{h_3h_4} + \frac{1}{2}S^2_{k-\frac{1}{2}}\nu_l\frac{\Delta t}{h_3^2} + \frac{1}{2}\left|\left(\rho\vartheta S_{k-\frac{1}{2}}\nu_l\right)\right|\frac{\Delta t}{h_3h_4}\right. \\
& + \frac{1}{2}\left|\left(\rho\vartheta S_k\nu_{l+\frac{1}{2}}\right)\right|\frac{\Delta t}{h_3h_4} + \frac{1}{2}\vartheta_k^2\nu_{l+\frac{1}{2}}\frac{\Delta t}{h_4^2} + \frac{1}{2}\left|\left(\rho\vartheta S_k\nu_{l-\frac{1}{2}}\right)\right|\frac{\Delta t}{h_3h_4} + \frac{1}{2}\vartheta_k^2\nu_{l-\frac{1}{2}}\frac{\Delta t}{h_4^2} \\
& + r|(\beta_j)|\frac{\Delta t}{h_2} + \left|\left(S_k\nu_l + \frac{1}{2}\rho\vartheta S_k - \mu S_k\right)\right|\frac{\Delta t}{h_3} + \left|\frac{1}{2}\rho\vartheta\nu_l + \frac{1}{2}\vartheta^2 - \xi(\eta-\nu_l)\right|\frac{\Delta t}{h_4} \\
& + m_1(n,i,j,k,l)\frac{\Delta t}{h_1} + m_1(n,i,j,k,l)(1+\theta)S_k\frac{\Delta t}{h_2} \\
& \left. + n_1(n,i,j,k,l)\frac{\Delta t}{h_1} + n_1(n,i,j,k,l)(1-\theta)S_k\frac{\Delta t}{h_2}\right] \\
& - V^n_{(i+1)jkl}\left[m_1(n,i,j,k,l)\frac{\Delta t}{h_1}\right] - V^n_{(i-1)jkl}\left[n_1(n,i,j,k,l)\frac{\Delta t}{h_1}\right] \\
& - V^n_{i(j+1)kl}\left[[r(\beta_j)^+ + n_1(n,i,j,k,l)(1-\theta)S_k]\frac{\Delta t}{h_2}\right] \\
& - V^n_{i(j-1)kl}\left[[r|(\beta_j)^-| + m_1(n,i,j,k,l)(1+\theta)S_k]\frac{\Delta t}{h_2}\right] \\
& - V^n_{ij(k+1)l}\left[\frac{1}{2}S^2_{k+\frac{1}{2}}\nu_l\frac{\Delta t}{h_3^2} + \frac{1}{2}\left(\rho\vartheta S_k\nu_{l+\frac{1}{2}}\right)^+\frac{\Delta t}{h_3h_4} + \frac{1}{2}\left|\left(\rho\vartheta S_k\nu_{l-\frac{1}{2}}\right)^-\right|\frac{\Delta t}{h_3h_4}\right. \\
& \left. + \left|\left(S_k\nu_l + \frac{1}{2}\rho\vartheta S_k - \mu S_k\right)^-\right|\frac{\Delta t}{h_3}\right] \\
& - V^n_{ij(k-1)l}\left[\frac{1}{2}S^2_{k-\frac{1}{2}}\nu_l\frac{\Delta t}{h_3^2} + \frac{1}{2}\left|\left(\rho\vartheta S_k\nu_{l+\frac{1}{2}}\right)^-\right|\frac{\Delta t}{h_3h_4} + \frac{1}{2}\left(\rho\vartheta S_k\nu_{l-\frac{1}{2}}\right)^+\frac{\Delta t}{h_3h_4}\right. \\
& \left. + \left(S_k\nu_l + \frac{1}{2}\rho\vartheta S_k - \mu S_k\right)^+\frac{\Delta t}{h_3}\right] \\
& - V^n_{ijk(l+1)}\left[\frac{1}{2}\left(\rho\vartheta S_{k+\frac{1}{2}}\nu_l\right)^+\frac{\Delta t}{h_3h_4} + \frac{1}{2}\left|\left(\rho\vartheta S_{k-\frac{1}{2}}\nu_l\right)^-\right|\frac{\Delta t}{h_3h_4} + \frac{1}{2}\vartheta_k^2\nu_{l+\frac{1}{2}}\frac{\Delta t}{h_4^2}\right. \\
& \left. + \left|\left(\frac{1}{2}\rho\vartheta\nu_l + \frac{1}{2}\vartheta^2 - \xi(\eta-\nu_l)\right)^-\right|\frac{\Delta t}{h_4}\right] \\
& - V^n_{ijk(l-1)}\left[\frac{1}{2}\left|\left(\rho\vartheta S_{k+\frac{1}{2}}\nu_l\right)^-\right|\frac{\Delta t}{h_3h_4} + \frac{1}{2}\left(\rho\vartheta S_{k-\frac{1}{2}}\nu_l\right)^+\frac{\Delta t}{h_3h_4} + \frac{1}{2}\vartheta_k^2\nu_{l-\frac{1}{2}}\frac{\Delta t}{h_4}\right. \\
& \left. + \left(\frac{1}{2}\rho\vartheta\nu_l + \frac{1}{2}\vartheta^2 - \xi(\eta-\nu_l)\right)^+\frac{\Delta t}{h_4}\right].
\end{aligned} \tag{38}
$$

Noting that the right-hand side of (38) contains no more than nine non-zero terms, the coefficient matrix of the system is septa-diagonal. Introduce an index transformation $q = q(i,j,k,l)$ for $i = 1,\ldots,E-1, j = 1,\ldots,M-1, k = 1,\ldots,P-1$ and $l = 1,..Z-1$ such that all the interior nodes of the mesh, i.e., those having indices in $G_h$, are re-ordered consecutively in such a way that $q(1,1,1,1) = 1, q(2,1,1,1) = 2,\ldots,q(E-1,M-1,P-1,Z-1) = (E-1)\times(M-1)\times(P-1)\times(Z-1) =: Q$. Let

$$
w^n_2(q(i,j,k,l)) = m_1(n,i,j,k,l)\frac{\Delta t}{h_1}, \qquad w^n_3(q(i,j,k,l)) = n_1(n,i,j,k,l)\frac{\Delta t}{h_1}
$$

DISTRIBUTION A. Approved for public release: distribution unlimited.

$$w_4^n(q(i,j,k,l)) = [r(\beta_j)^+ + n_1(n,i,j,k,l)(1-\theta)S_k]\frac{\Delta t}{h_2},$$
$$w_5^n(q(i,j,k,l)) = [r|(\beta_j)^-| + m_1(n,i,j,k,l)(1+\theta)S_k]\frac{\Delta t}{h_2},$$

$$\begin{aligned} w_6^n(q(i,j,k,l)) &= \frac{1}{2}S_{k+\frac{1}{2}}^2\nu_l\frac{\Delta t}{h_3^2} + \frac{1}{2}\left(\rho\vartheta S_k\nu_{l+\frac{1}{2}}\right)^+\frac{\Delta t}{h_3h_4} + \frac{1}{2}\left|\left(\rho\vartheta S_k\nu_{l-\frac{1}{2}}\right)^-\right|\frac{\Delta t}{h_3h_4} \\ &\quad + \left|\left(S_k\nu_l + \frac{1}{2}\rho\vartheta S_k - \mu S_k\right)^-\right|\frac{\Delta t}{h_3} \\ w_7^n(q(i,j,k,l)) &= \frac{1}{2}S_{k-\frac{1}{2}}^2\nu_l\frac{\Delta t}{h_3^2} + \frac{1}{2}\left|\left(\rho\vartheta S_k\nu_{l+\frac{1}{2}}\right)^-\right|\frac{\Delta t}{h_3h_4} + \frac{1}{2}\left(\rho\vartheta S_k\nu_{l-\frac{1}{2}}\right)^+\frac{\Delta t}{h_3h_4} \\ &\quad + \left(S_k\nu_l + \frac{1}{2}\rho\vartheta S_k - \mu S_k\right)^+\frac{\Delta t}{h_3}, \\ w_8^n(q(i,j,k,l)) &= \frac{1}{2}\left(\rho\vartheta S_{k+\frac{1}{2}}\nu_l\right)^+\frac{\Delta t}{h_3h_4} + \frac{1}{2}\left|\left(\rho\vartheta S_{k-\frac{1}{2}}\nu_l\right)^-\right|\frac{\Delta t}{h_3h_4} + \frac{1}{2}\vartheta_k^2\nu_{l+\frac{1}{2}}\frac{\Delta t}{h_4^2} \\ &\quad + \left|\left(\frac{1}{2}\rho\vartheta\nu_l + \frac{1}{2}\vartheta^2 - \xi(\eta-\nu_l)\right)^-\right|\frac{\Delta t}{h_4}, \\ w_9^n(q(i,j,k,l)) &= \frac{1}{2}\left|\left(\rho\vartheta S_{k+\frac{1}{2}}\nu_l\right)^-\right|\frac{\Delta t}{h_3h_4} + \frac{1}{2}\left(\rho\vartheta S_{k-\frac{1}{2}}\nu_l\right)^+\frac{\Delta t}{h_3h_4} + \frac{1}{2}\vartheta_k^2\nu_{l-\frac{1}{2}}\frac{\Delta t}{h_4} \\ &\quad + \left(\frac{1}{2}\rho\vartheta\nu_l + \frac{1}{2}\vartheta^2 - \xi(\eta-\nu_l)\right)^+\frac{\Delta t}{h_4}. \end{aligned}$$

It is clear that

$$w_l^n(q(i,j,k,l)) \geq 0 \quad l = 2,3,\ldots,9, \tag{39}$$
$$w_1^n(q(i,j,k,l)) = 1 + \sum_{l=2}^{9} w_l^n(q(i,j,k,l)) \geq 1. \tag{40}$$

Using the above notation, we can write (38) as the following form:

$$\begin{aligned} V_{ijkl}^{n+1} &= V_{ijkl}^n w_1^n(q(i,j,k,l)) - V_{(i+1)jkl}^n w_2^n(q(i,j,k,l)) - V_{(i-1)jkl}^n w_3^n(q(i,j,k,l)) \\ &\quad - V_{i(j+1)kl}^n w_4^n(q(i,j,k,l)) - V_{i(j-1)k,l}^n w_5^n(q(i,j,k,l)) - V_{ij(k+1)l}^n w_6^n(q(i,j,k,l)) \\ &\quad - V_{ij(k-1)l}^n w_7^n(q(i,j,k,l)) - V_{ijk(l+1)}^n w_8^n(q(i,j,k,l)) - V_{ijk(l-1)}^n w_9^n(q(i,j,k,l)) \end{aligned} \tag{41}$$

for $(i,j,k,l) \in G_h$.

Casting the terms associated with Dirichlet boundary points to the LHS and swapping the LHS and RHS of the resulting system, we see that (41) can be rewritten in the following matrix form:

$$A^n V^n = b^n + c^{n+1} \tag{42}$$

for $n = N-1, N-2, \ldots, 0$, where $A^n = (a_{pq}^n)_{p,q=1}^Q$ is a septa-diagonal matrix with the non-zero entries given by

$$a_{qq}^n = w_1^n(q), \qquad a_{q,q+1}^n = -w_2^n(q), \qquad a_{q,q+(E-1)}^n = -w_4^n(q), \tag{43}$$
$$a_{q,q+(E-1)\times(M-1)}^n = -w_6^n(q), \qquad a_{q,q+(E-1)\times(M-1)\times(P-1)}^n = -w_8^n(q), \tag{44}$$
$$a_{q,q-1}^n = -w_3^n(q), \qquad a_{q,q-(E-1)}^n = -w_5^n(q), \tag{45}$$
$$a_{q,q-(E-1)\times(M-2)}^n = -w_7^n(q), \qquad a_{q,q-(E-1)\times(M-1)\times(P-1)}^n = -w_9^n(q). \tag{46}$$

for $q = 1,2,\ldots,Q$, $b^n$ and $c^{n+1}$ are $Q \times 1$ column vectors representing, respectively, the contribution from the Dirichlet boundary conditions at the $n$th time step and the left-hand side of (41) involving the approximate solution at the $(n+1)$th time step, and $V^n$ denotes the unknown vector. This is a $Q \times Q$ linear system in $V^n$. From (39) and (40), (43)–(46) it is easy to see that $A^n$ is diagonally dominant, irreducible and has positive diagonal and non-positive off-diagonal elements. Thus, $A^n$ is an $M$-matrix (cf., for example, [30]) and is non-singular. Therefore, we conclude that there exists a unique solution to (42)/(32) with given $m_1(n,i,j,k,l) \in [0,\lambda]$ and $n_1(n,i,j,k,l) \in [0,\lambda]$.

DISTRIBUTION A. Approved for public release: distribution unlimited.

We next show that for any $\Delta > (0, 0)$, the solution is uniformly bounded. Our strategy is to prove that if the terminal conditions satisfy

$$\max_{(i,j,k,l)\in G_h} |V_{ijkl}^N| < C < +\infty$$

for some positive constant $C$, then

$$\max_{(i,j,k,l)\in G_h} |V_{ijkl}^n| < C, \quad n = 1, 2, .., N-1. \tag{47}$$

By (35)–(36) and (6), we have

$$\max_{(i,j,k,l)\in G_h} |V_{ijkl}^N| < 1.$$

Let $\max_{(i,j,k,l)\in G_h} |V_{ijkl}^n| < 1$ hold for an $n \le N$, we will show that $\max_{(i,j,k,l)\in G_h} |V_{ijkl}^{n-1}| < 1$ by contradiction.

Suppose that $\max_{(i,j,k,l)\in G_h} |V_{ijkl}^{n-1}| \ge 1$. Then, there exists an index triple $(i_0, j_0, k_0, l_0) \in G_h$, such that

$$|V_{i_0j_0k_0l_0}^{n-1}| \ge 1 \quad \text{and} \quad |V_{ijkl}^{n-1}| \le |V_{i_0j_0k_0l_0}^{n-1}|, \quad \forall (i, j, k, l) \in G_h.$$

Combining this with (41), we have

$$\begin{aligned}
|V_{i_0j_0k_0l_0}^n| \ge & \left| V_{i_0j_0k_0l_0}^{n-1} \left(1 + \sum_{p=2}^{9} w_p^{n-1}(q_0)\right) \right| - |V_{(i_0+1)j_0k_0l_0}^{n-1} w_2^{n-1}(q_0)| - |V_{(i_0-1)j_0k_0l_0}^{n-1} w_3^{n-1}(q_0)| - |V_{i_0(j_0+1)k_0l_0}^{n-1} w_4^{n-1}(q_0)| \\
& - |V_{i_0(j_0-1)k_0l_0}^{n-1} w_5^{n-1}(q_0)| - |V_{i_0j_0(k_0+1)l_0}^{n-1} w_6^{n-1}(q_0)| - |V_{i_0j_0(k_0-1)l_0}^{n-1} w_7^{n-1}(q_0)| \\
& - |V_{i_0j_0k_0(l_0+1)}^{n-1} w_8^{n-1}(q_0)| - |V_{i_0j_0k_0(l_0-1)}^{n-1} w_9^{n-1}(q_0)| \\
= & |V_{i_0j_0k_0l_0}^{n-1}| + |w_2^{n-1}(q_0)|(|V_{i_0j_0k_0l_0}^{n-1}| - |V_{(i_0+1)j_0k_0l_0}^{n-1}|) + |w_3^{n-1}(q_0)|(|V_{i_0j_0k_0l_0}^{n-1}| - |V_{(i_0-1)j_0k_0l_0}^{n-1}|) \\
& + |w_4^{n-1}(q_0)|(|V_{i_0j_0k_0l_0}^{n-1}| - |V_{i_0(j_0+1)k_0l_0}^{n-1}|) + |w_5^{n-1}(q_0)|(|V_{i_0j_0k_0l_0}^{n-1}| - |V_{i_0(j_0-1)k_0l_0}^{n-1}|) \\
& + |w_6^{n-1}(q_0)|(|V_{i_0j_0k_0l_0}^{n-1}| - |V_{i_0j_0(k_0+1)l_0}^{n-1}|) + |w_7^{n-1}(q_0)|(|V_{i_0j_0k_0l_0}^{n-1}| - |V_{i_0j_0(k_0-1)l_0}^{n-1}|) \\
& + |w_8^{n-1}(q_0)|(|V_{i_0j_0k_0l_0}^{n-1}| - |V_{i_0j_0k_0(l_0+1)}^{n-1}|) + |w_9^{n-1}(q_0)|(|V_{i_0j_0k_0l_0}^{n-1}| - |V_{i_0j_0k_0(l_0-1)}^{n-1}|) \\
\ge & |V_{i_0j_0k_0l_0}^{n-1}| \\
\ge & 1,
\end{aligned}$$

where $q_0 = q(i_0, j_0, k_0, l_0)$. Clearly, this contradicts our assumption that $\max_{(i,j,k,l)\in G_h} |V_{ijkl}^n| < 1$. Thus, we have $\max_{(i,j,k,l)\in G_h} |V_{ijkl}^{n-1}| < 1$. By the mathematical induction principle, (47) holds and the theorem is proved. □

To conclude this section, we comment since the system matrix of (32) is an $M$-matrix by Theorem 3.2 the discretization is monotone which guarantees that the solution to (32) is non-negative since the boundary conditions (35)–(36) are non-negative.

## 4. Decoupling algorithm and its convergence

In this section we present a decoupling algorithm for solving the nonlinear system (42)/(32)–(34). We first rewrite (37), which is equivalent to (42) and (32), as the following form:

$$\mathcal{L}_1^{\Delta} V_{ijkl}^n + \min_{\bar{m}\in[0,\lambda]} \bar{m}\mathcal{L}_2^{\Delta} V_{ijkl}^n + \min_{\bar{n}\in[0,\lambda]} \bar{n}\mathcal{L}_3^{\Delta} V_{ijkl}^n = 0 \tag{48}$$

for $n = N-1, N-2, \ldots, 0$ and $(i, j, k, l) \in G_h$, where $\Delta = (\Delta t, h)$ as defined previously and

$$\begin{aligned}
\mathcal{L}_1^{\Delta} V_{ijkl}^n := & -\frac{V_{ijkl}^{n+1} - V_{ijkl}^n}{\Delta t} h_1h_2h_3h_4 - \frac{1}{2} S_{k+\frac{1}{2}}^2 \nu_l \frac{V_{ij(k+1)l}^n - V_{ijkl}^n}{h_3} h_1h_2h_4 \\
& - \frac{1}{2}\left(\rho\vartheta S_{k+\frac{1}{2}}\nu_l\right)^+ \frac{V_{ijk(l+1)}^n - V_{ijkl}^n}{h_4} \times h_1h_2h_4 + \frac{1}{2}\left|\left(\rho\vartheta S_{k+\frac{1}{2}}\nu_l\right)^-\right| \frac{V_{ijkl}^n - V_{ijk(l-1)}^n}{h_4} h_1h_2h_4 \\
& + \frac{1}{2} S_{k-\frac{1}{2}}^2 \nu_l \frac{V_{ijkl}^n - V_{ij(k-1)l}^n}{h_3} \times h_1h_2h_4 + \frac{1}{2}\left(\rho\vartheta S_{k-\frac{1}{2}}\nu_l\right)^+ \frac{V_{ijkl}^n - V_{ijk(l-1)}^n}{h_4} h_1h_2h_4 \\
& - \frac{1}{2}\left|\left(\rho\vartheta S_{k-\frac{1}{2}}\nu_l\right)^-\right| \frac{V_{ijk(l+1)}^n - V_{ijkl}^n}{h_4} h_1h_2h_4 - \frac{1}{2}\left(\rho\vartheta S_k \nu_{l+\frac{1}{2}}\right)^+ \frac{V_{ij(k+1)l}^n - V_{ijkl}^n}{h_3} h_1h_2h_3
\end{aligned}$$

DISTRIBUTION A. Approved for public release: distribution unlimited.

$$
\begin{aligned}
&+\frac{1}{2}\left|\left(\rho\vartheta S_k\nu_{l+\frac{1}{2}}\right)^-\right|\frac{V^n_{ijkl}-V^n_{ij(k-1)l}}{h_3}h_1h_2h_3-\frac{1}{2}\vartheta_k^2\nu_{l+\frac{1}{2}}\frac{V^n_{ijk(l+1)}-V^n_{ijkl}}{h_4}h_1h_2h_3\\
&+\frac{1}{2}\left(\rho\vartheta S_k\nu_{l-\frac{1}{2}}\right)^+\frac{V^n_{ijkl}-V^n_{ij(k-1)l}}{h_3}h_1h_2h_3-\frac{1}{2}\left|\left(\rho\vartheta S_k\nu_{l-\frac{1}{2}}\right)^-\right|\frac{V^n_{ij(k+1)l}-V^n_{ijkl}}{h_3}h_1h_2h_3\\
&+\frac{1}{2}\vartheta_k^2\nu_{l-\frac{1}{2}}\frac{V^n_{ijkl}-V^n_{ijk(l-1)}}{h_4}h_1h_2h_3-r(\beta_j)^+\frac{V^n_{i(j+1)kl}-V^n_{ijkl}}{h_2}h_1h_2h_3h_4\\
&+r|(\beta_j)^-|\frac{V^n_{ijkl}-V^n_{i(j-1)kl}}{h_2}h_1h_2h_3h_4+\left(S_k\nu_l+\frac{1}{2}\rho\vartheta S_k-\mu S_k\right)^+\frac{V^n_{ijkl}-V^n_{ij(k-1)l}}{h_3}h_1h_2h_3h_4\\
&-\left|\left(S_k\nu_l+\frac{1}{2}\rho\vartheta S_k-\mu S_k\right)^-\right|\frac{V^n_{ij(k+1)l}-V^n_{ijkl}}{h_3}h_1h_2h_3h_4\\
&+\left(\frac{1}{2}\rho\vartheta\nu_l+\frac{1}{2}\vartheta^2-\xi(\eta-\nu_l)\right)^+\frac{V^n_{ijkl}-V^n_{ijk(l-1)}}{h_4}h_1h_2h_3h_4\\
&-\left|\left(\frac{1}{2}\rho\vartheta\nu_l+\frac{1}{2}\vartheta^2-\xi(\eta-\nu_l)\right)^-\right|\frac{V^n_{ijk(l+1)}-V^n_{ijkl}}{h_4}h_1h_2h_3h_4
\end{aligned}
\tag{49}
$$

$$
\mathcal{L}_2^{\Delta}V^n_{ijkl}:=\left(-\frac{V^n_{(i+1)jkl}-V^n_{ijkl}}{h_1}+(1+\theta)S_k\frac{V^n_{ijkl}-V^n_{i(j-1)kl}}{h_2}\right)h_1h_2h_3h_4 \tag{50}
$$

$$
\mathcal{L}_3^{\Delta}V^n_{ijkl}:=\left(\frac{V^n_{ijkl}-V^n_{(i-1)jkl}}{h_1}-(1-\theta)S_k\frac{V^n_{i(j+1)kl}-V^n_{ijkl}}{h_2}\right)h_1h_2h_3h_4. \tag{51}
$$

We propose the following algorithm for (42).

**Algorithm D.** 1. Initialize $V^N_{ijkl}$ for all $(i,j,k,l)\in\bar{G}_h$ using the terminal and boundary conditions (35)–(36) and let $n=N-1$.

2. Let $V^{n,0}_{ijkl}=V^{n+1}_{ijkl}$ for all $(i,j,k,l)\in\bar{G}_h$ and evaluate

$$
\begin{aligned}
m_1^0(n,i,j,k,l)&=\arg\left(\min_{\bar{m}\in[0,\lambda]}\bar{m}\mathcal{L}_2^{\Delta}V^{n+1}_{ijkl}\right),\\
n_1^0(n,i,j,k,l)&=\arg\left(\min_{\bar{n}\in[0,\lambda]}\bar{n}\mathcal{L}_3^{\Delta}V^{n+1}_{ijkl}\right)
\end{aligned}
$$

for any $(i,j,k,l)\in G_h$.

3. For a given tolerance $\varepsilon>0$, set $p=0$.

4. Solve, $\forall(i,j,k,l)\in G_h$, the following system along with the boundary conditions (36) for $\{V^{n,p+1}_{ijkl}\}_{(i,j,k,l)\in G_h}$:

$$
\mathcal{L}_1^{\Delta}V^{n,p+1}_{ijkl}+m^p(n,i,j,k,l)\mathcal{L}_2^{\Delta}V^{n,p+1}_{ijkl}+n^p(n,i,j,k,l)\mathcal{L}_3^{\Delta}V^{n,p+1}_{ijkl}=0, \tag{52}
$$

where, when applied to $V^{n,p+1}_{ijkl}$, the finite difference operator is

$$
D_tV^{n,p+1}_{ijkl}=\frac{V^{n+1}_{ijkl}-V^{n,p+1}_{ijkl}}{\Delta t}. \tag{53}
$$

5. Evaluate, for all $(i,j,k,l)\in G_h$,

$$
m_1^{p+1}(n,i,j,k,l)=\arg\left(\min_{\bar{m}\in[0,\lambda]}\bar{m}\mathcal{L}_2^{\Delta}V^{n,p+1}_{ijkl}\right), \tag{54}
$$

$$
n_1^{p+1}(n,i,j,k,l)=\arg\left(\min_{\bar{n}\in[0,\lambda]}\bar{n}\mathcal{L}_3^{\Delta}V^{n,p+1}_{ijk}\right). \tag{55}
$$

6. If $\max_{(i,j,k,l)\in G_h}|V^{n,p+1}(i,j,k,l)-V^{n,p}(i,j,k,l)|\geq\varepsilon$, set $p=p+1$ and goto Step 4. Otherwise, goto Step 7.

7. Set $V^n_{ijkl}=V^{n,p+1}_{ijkl}$ for $(i,j,k,l)\in\bar{G}_h$ and $m_1(n,i,j,k,l)=m_1^{p+1}(n,i,j,k,l)$, $n_1(n,i,j,k,l)=n_1^{p+1}(n,i,j,k,l)$ for $(i,j,k,l)\in G_h$. If $n=0$, stop. Otherwise, let $n=n-1$ and goto Step 2.

Clearly, in Algorithm D, the nonlinear system (32)–(34) is decoupled so that in each iteration we only solve the linear system (52).

Using the notation used in the proof of Theorem 3.2, we let $V^{n,p+1}$ denote the solution to (52). Then, the convergence of the iterative algorithm in Algorithm D is given in the following theorem.

DISTRIBUTION A. Approved for public release: distribution unlimited.

**Theorem 4.1.** *The iterative scheme (52)–(55) generates a sequence* $\{V^{n,p}\}_{p=0}^{\infty}$ *that converges to the solution of (48) with the terminal and boundary conditions (35)–(36).*

**Proof.** We will use the notation in Algorithm D. To prove this theorem, we first show that the sequence $\{V^{n,p}\}_{p=0}^{\infty}$ generated by the iterative method is monotonically increasing, i.e., $V^{n,p} \le V^{n,p+1}$ for $p \ge 1$.

From (52) we have

$$\mathcal{L}_1^{\Delta} V_{ijkl}^{n,p} + m_1^{p-1}(n,i,j,k,l)\mathcal{L}_2^{\Delta} V_{ijkl}^{n,p} + n_1^{p-1}(n,i,j,k,l)\mathcal{L}_3^{\Delta} V_{ijkl}^{n,p} = 0, \quad (i,j,k,l) \in G_h$$

for $p = 1, 2, \ldots$. This can be rewritten as

$$\begin{aligned}
&\mathcal{L}_1^{\Delta} V_{ijkl}^{n,p} + m_1^{p}(n,i,j,k,l)\mathcal{L}_2^{\Delta} V_{ijkl}^{n,p} + n_1^{p}(n,i,j,k,l)\mathcal{L}_3^{\Delta} V_{ijkl}^{n,p} \\
&\quad = [m_1^{p}(n,i,j,k,l) - m_1^{p-1}(n,i,j,k,l)]\mathcal{L}_2^{\Delta} V_{ijkl}^{n,p} \\
&\qquad + [n_1^{p}(n,i,j,k,l) - n_1^{p-1}(n,i,j,k,l)]\mathcal{L}_3^{\Delta} V_{ijkl}^{n,p} \le 0, \quad \forall (i,j,k,l) \in G_h,
\end{aligned} \tag{56}$$

since, by (54) and (55),

$$m_1^{p}(n,i,j,k,l) = \arg\left(\min_{\bar{m}\in[0,\lambda]} \bar{m}\mathcal{L}_2^{\Delta} V_{ijkl}^{n,p}\right), \qquad n_1^{p}(n,i,j,k,l) = \arg\left(\min_{\bar{n}\in[0,\lambda]} \bar{n}\mathcal{L}_3^{\Delta} V_{ijkl}^{n,p}\right).$$

Note that both (52) and (56) have the same boundary conditions, i.e., $V_{ijkl}^{n,p} = V_{ijkl}^{n,p+1}$ for any $(i,j,k,l) \in \partial G_h$. Thus, using the notation in the proof of Theorem 3.2, we may write (52) and (56) as the following respective matrix forms similar to (42):

$$A^{n,p}V^{n,p+1} = b^n + c^{n+1} \quad \text{and} \quad A^{n,p}V^{n,p} \le b^n + c^{n+1},$$

where $A^{n,p}$, $b^n$ and $c^{n+1}$ are as defined in (42) with $A^{n,p}$ an $M$-matrix. Therefore, we have

$$A^{n,p}(V^{n,p+1} - V^{n,p}) \ge 0.$$

Since $A^{n,p}$ is an $M$-matrix, we have

$$V^{n,p+1} - V^{n,p} \ge 0.$$

Therefore, the monotonicity of iteration process is proved.

From Theorem 3.2 we have that $V^{n,p}$ is bounded for any $p = 0, 1, 2, \ldots$. Combining the monotonicity and boundedness of $V^{n,p}$ we see that $V^{n,p}$ is convergent. Finally, from the construction of (52) and (55) it is obvious that $V_{ijkl}^{n,p}$, $m_1^{p}(n,i,j,k,l)$ and $n_1^{p}(n,i,j,k,l)$ solve (48) when $p \to \infty$. Thus, we have proved the theorem. □

## 5. Numerical results

In this section, we use the scheme (32)–(36) to calculate the value functions $V^i(i = 0, b, w)$ and present the computed reservation purchase and write prices of a European call option by using the utility function in (6). Note that using this utility function can eliminate the bond account variable, $\beta$, in (12) by a transformation. However, in this paper, we will implement our schemes without eliminating $\beta$ to demonstrate our algorithm can also be used for other types of utility functions.

We now illustrate the performance and usefulness of the scheme using the following test example:

**Test Example**: Reservation purchase and write prices of a European call option with expiry date $T = 0.6$, the initial price of the underlying stock $S_0 = 1.6$, the initial value of volatility $v_0 = 1.6$, the risk aversion parameter $\gamma = 1$ with various values of $\alpha_0$, $\beta_0$ and strike price $K$. Other parameters are: $r = 0.05$, $\mu = 0.1$, $\theta = 0.005$, $\xi = 5$, $\eta = 0.16$, $\rho = 0.1$ and $\vartheta = 0.9$

To solve the problem, we choose $\underline{L}_\alpha = \underline{L}_\beta = 1$, $\bar{L}_\alpha = 3$, $\bar{L}_\beta = 5$ and $\bar{L}_S = \bar{L}_v = 3$ in (24). The mesh and penalty parameters are chosen to be $E = 20$, $M = 30$, $P = Z = 15$, $N = 15$ and $\lambda = 1000$. Other parameters are $\Delta t = 0.04$, $h = 0.2$. Using the numerical solution on this mesh, we examine the changes of reservation purchase and write prices with respect to different variables.

We first examine the influence of $\beta_0$ on reservation prices. For $\alpha_0 = 2$ and $K = 1.6$, we compute the reservation purchase and write prices with various values of $\beta_0$ and the computed results are plotted in Fig. 1. It is clear from Fig. 1 that the initial holding in the bond does not affect both the reservation purchase and write prices. This coincides with the results in [19]. As mentioned before, using an exponential utility function can eliminate the bond account variable $\beta$ by a transformation. Thus, the resulting option prices are independent of $\beta_0$.

To examine the influence of $\alpha_0$ on the reservation prices, we compute the purchase and write prices for $\alpha_0 = 1.6, 2, 2.4$ using our numerical method. Since the results are independent of $\beta_0$, we only plot the computed purchase and write prices $P_b$ and $P_w$ against $\alpha_0$ in Fig. 2. From Fig. 2 we have the following observations:

1. The write prices first decrease as $\alpha_0$ increase and then tend to be stable, and
2. the reservation write price is always higher than the purchase price.

DISTRIBUTION A. Approved for public release: distribution unlimited.

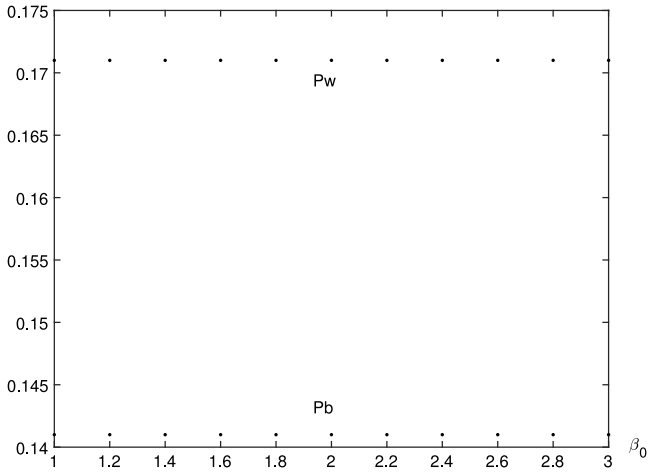



**Fig. 1.** Computed reservation prices for $\alpha_0 = 2$, $K = 1.6$ and $\rho = 0.1$.

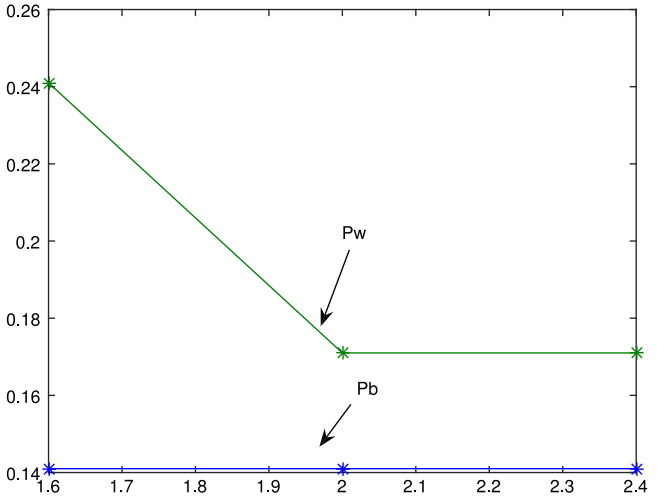



**Fig. 2.** Computed reservation prices for $\alpha_0 = 1.6$, 2 and 2.4.

These observations are financially correct because of the following reasons. It is known that the utility based option pricing approach values an option from both buyers' and sellers' perspectives, which leads to two prices, i.e., reservation write and purchase prices. The reservation write and purchase prices are respectively the prices at which the investor is willing to sell and the investor is willing to purchase the option. Clearly, the write price is greater than the purchase price as the price that the seller wants to receive is always higher than that the buyer wants to pay. Also, according to the theory of supply and demand, the more stock a writer (respectively purchaser) holds, the more (respectively less) he/she wants to sell (respectively purchase) the stock and therefore he/she will reduce the option price. However, when the price reaches a certain level, it will not be reduced further.

Finally, we consider the influence of the strike price $K$ on the reservation prices. To achieve this, we assume that $\alpha_0 = 2$ and compute the purchase and write prices for $K = 1$, 1.6 and 2.4. Again, since the computed prices are independent of $\beta_0$, we plot them against $K$ in Fig. 3 from which it is easily seen that both the purchase and write prices decrease as $K$ increases. The explanation for this phenomenon is as follows. If $S_T > K$, the option holder (buyer) will exercise the option at the expiry date $T$. Thus, the buyer will earn $S_T - K$ and the writer will lost the same amount. Since the gain/loss $S_T - K$ is a decreasing functions of $K$, when $K$ increases, both buyer and writer will reduce the option price.

We remark that the original problem is defined on an infinite domain and does not have any Dirichlet boundary conditions. In the paper, we define an artificial (homogeneous) Dirichlet boundary condition on each of the boundary segments as the exact one is unknown. The computational errors caused by the above artificial boundary condition are essentially located in the boundary layer, as shown in [29]. Thus, in the numerical results presented above, we only plot the computed values at the mesh points which are some distance away from the boundary segments.

**DISTRIBUTION A. Approved for public release: distribution unlimited.**

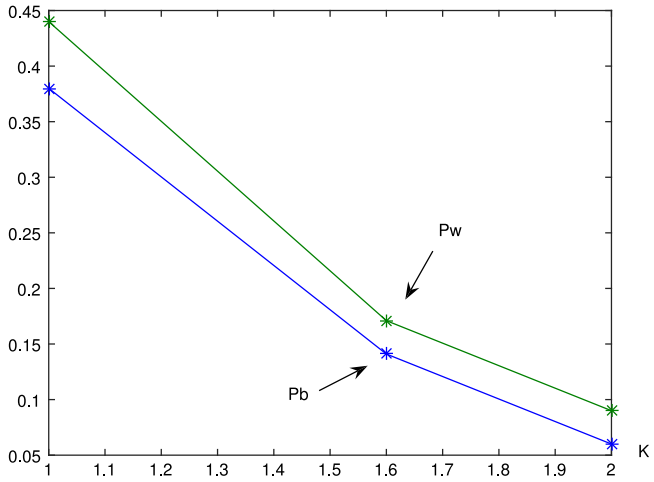



**Fig. 3.** Computed reservation prices for $K = 1$, 1.6 and 2.4.

## 6. Conclusion

In this paper we propose a penalty method combined with a finite volume scheme to solve the HJB equation in 4 spatial dimensions governing the reservation purchase and write prices of a European call option with proportional transaction costs and stochastic volatility. This scheme has the merits that it is easy to implement and the resulting system matrix is an $M$-matrix. The latter guarantees that numerical solutions from discretization method are always non-negative when the boundary and payoff conditions are non-negative. The numerical results showed that the method is able to solve problems of practical significance.

## Acknowledgement

This work is partially supported by the AOARD Project #15IOA095 from the US Air Force.

## References

[1] F. Black, M. Scholes, The pricing of options and corporate liabilities, J. Polit. Econ. 81 (1973) 637–659.
[2] P.P. Boyle, T. Vorst, Option replication in discrete time with transaction costs, J. Finance 47 (1) (1992) 271–293.
[3] M.H.A. Davis, V.G. Panas, Z. Zariphopoulou, European option pricing with transaction costs, SIAM J. Control Optim. 31 (2) (1993) 470–493.
[4] C. Edirisinghe, V. Naik, R. Uppal, Optimal replication of options with transaction costs and trading restrictions, J. Finan. Quant. Anal. 28 (1993) 117–138.
[5] S. Figlewski, Options arbitrage in imperfect markets, J. Finance 44 (5) (1989) 1289–1311.
[6] S.D. Hodges, A. Neuberger, Optimal replication of contingent claims under transaction costs, Rev. Future Mark. 8 (1989) 222–239.
[7] H.E. Leland, Option pricing and replication with transaction costs, J. Finance 40 (1985) 1283–1301.
[8] K.B. Toft, On the Mean-Variance Tradeoff in option replication with transaction costs, J. Finan. Quant. Anal. 31 (2) (1996) 233–262.
[9] A. Damgaard, Utility based option evaluation with proportional transaction costs, J. Econom. Dynam. Control 27 (2003) 667–700.
[10] L. Clewlow, S. Hodge, Optimal Delta-Hedging under transaction costs, J. Econom. Dynam. Control 21 (1997) 1353–1376.
[11] A. Damgaard, Computation of reservation prices of options with proportional transaction costs, J. Econ. Dyn. Control 30 (2006) 415–444.
[12] M.H.A. Davis, T. Zariphopoulou, in: M.H.A. Davis, et al. (Eds.), American Options and Transaction Fees, in: Mathemtical Finance, Springer-Verlag, New York, 1995.
[13] M. Monoyios, Option pricing with transaction costs using a Markov chain approximation, J. Econom. Dynam. Control 28 (2004) 889–913.
[14] V.I. Zakamouline, European option pricing and hedging with both fixed and proportional transaction costs, J. Econ. Dyn. Control 30 (2006) 1–25.
[15] V.I. Zakamouline, American option pricing and exercising with transaction costs, J. Comput. Finance 8 (3) (2005) 81–115.
[16] W. Li, S. Wang, Penalty approach to the HJB equation arising in European stock option pricing with proportional transaction costs, J. Optim. Theory Appl. 143 (2009) 279–293.
[17] W. Li, S. Wang, A numerical method for pricing European options with proportional transaction costs, J. Global Optim. 60 (1) (2014) 59–78.
[18] W. Li, S. Wang, Pricing American options under proportional transaction costs using a penalty approach and a finite difference scheme, J. Ind. Manag. Optim. 9 (2) (2013) 365–389.
[19] R.E. Caflisch, G. Gambino, M. Sammartino, C. Sgarra, European option pricing with transaction costs and stochastic volatility: an asymptotic analysis, IMA J. Appl. Math. 80 (4) (2015) 981–1008.
[20] A. Cosso, D. Marazzina, C. Sgarra, American option valuation in a stochastic volatility model with transaction costs, Stochastics 87 (3) (2015) 518–536.
[21] S. Wang, X. Yang, K.L. Teo, A power penalty method for a complementarity problem arising from American option valuation, J. Optim. Theory Appl. 129 (2006) 227–254.
[22] S. Wang, A penalty method for a finite-dimensional obstacle problem with derivative constraints, Optim. Lett. 8 (2014) 1799–1811.
[23] S. Wang, A penalty approach to a discretized double obstacle problem with derivative constraints, J. Global Optim. 62 (2015) 775–790.
[24] W. Chen, S. Wang, A finite difference method for pricing European and American options under a geometric Levy process, J. Ind. Manag. Optim. 11 (2015) 241–264.
[25] D.C. Lesmana, S. Wang, Penalty approach to a nonlinear obstacle problem governing American put option valuation under transaction costs, Appl. Math. Comput. 251 (2015) 318–330.
[26] S. Wang, A novel fitted finite volume method for the Black–Scholes equation governing option pricing, IMA J. Numer. Anal. 24 (2004) 669–720.

**DISTRIBUTION A. Approved for public release: distribution unlimited.**

[27] C.-S. Huang, C.-H. Hung, S. Wang, A fitted finite volume method for the valuation of options on assets with stochastic volatilities, Computing 77 (2006) 297–320.
[28] S. Wang, S. Zhang, Z. Fang, A superconvergent fitted finite volume method for Black–Scholes equations governing European and American option valuation, Numer. Partial Differential Equations 31 (2015) 1190–1208.
[29] S. Richardson, S. Wang, The viscosity approximation to the Hamilton–Jacobi-Bellman equation in optimal feedback control: upper bounds for extended domains, J. Ind. Manag. Optim. 6 (2010) 161–175.
[30] R.S. Varga, Matrix Iterative Analysis, Prentice-Hall, Englewood Cliffs, New Jersey, 1962.

DISTRIBUTION A. Approved for public release: distribution unlimited.

Pacific Journal of Optimization
Vol. 12, No. 3, pp. 557–586, July 2016

# ANALYZING HUMAN PERIODIC WALKING AT DIFFERENT SPEEDS USING PARAMETERIZATION ENHANCING TRANSFORM IN DYNAMIC OPTIMIZATION

Meiyi Tan, Leslie S. Jennings and Song Wang*

**Abstract:** In this paper, we extend the human walking model proposed in [18] to improve periodic motion and to explore different walking speeds. We first propose the inclusion of additional constraints to better model the periodic motion of a human. We then introduce the use of Control Parameterization Enhancing Transform (CPET) technique within the model, along with the inclusion of velocity in the objective function, to allow different walking speeds to be used in the model. Numerical experiments, performed to show the superiority of this new model, show that periodic motion can be improved with the addition of periodicity constraints and that human walking motions can be replicated at faster or slower velocities, which is desirable in practice. The numerical results also show that the time interval in the single support and double support phases can be optimized using CPET for different walking speeds.

**Key words:** *optimal control of robot motion, dynamic optimization, biped locomotion, robot modelling and optimization.*

**Mathematics Subject Classification:** *49M30, 90C90, 93B40.*

## 1 Introduction

Walking, though classified as being the most basic human motion, is deemed to involve great complexities and thus should be given significant attention. It is considered one of the most complicated motions involving a series of complex continuous and discontinuous phases [22]. Modelling and optimization of human walking are key to robot design and application in industries such as manufacturing and health care. Human locomotion has attracted much attention from researchers and practitioners since 1970s and many of these results have been fundamental for advancement in the development of support for multiple movement disabilities [4]. The greater understanding of the mechanics behind walking can contribute to the ability to improve aids for people with locomotor disabilities, and the development of walking robots etc.

There has been a significant increase in research on human walking in the past decade. With the advancement of computing power, simulation of human walking governed by complex equations has now become possible. As demonstrated by existing studies, research based on modelling and simulation contributes more significantly to the study of human locomotion than empirical studies, since it is able to investigate in more detail the muscle

*S. Wang's work was partially supported by the AOARD Project #15IOA095 from the US Air Force.

ISSN 1348-9151 © 2016 Yokohama Publishers

DISTRIBUTION A. Approved for public release: distribution unlimited.

activities involved in the movement patterns, and can also lead to the accurate prediction of motion ([14, 15]). Amongst many research activities done on human walking, only a handful of them was on modelling and simulating the dynamics of human walking ([14, 16]).

The purpose of this study is to extend the model in [18] to one that improves periodic motion and allows different walking speeds in optimization. These are achieved by re-design of the constraints, introduction of velocity in the objective function and the use of the Control Parameterization Enhancing Transform (CPET) technique [21] in the optimal control model. In the current model, walking speeds are adjustable and can be computed so that an evaluation of the dynamics involved can be explored. The study provides insights to factors which are important in walking and can aid in future development of more realistic walking robots or aids for the locomotory handicapped.

Various algorithms are available to solve such problems and have been reviewed in optimization software guides [13]. However, many of these algorithms were not developed into general purpose software packages to solve optimal control problems that includes complex constraints. Software packages such as MATLAB or GAMS (which runs on MATLAB's platform) [3] may be capable of solving such problems, but are computationally slow. For these reasons, the MISER3 [8] optimal control software was chosen. The central idea behind the software MISER3 is the concept of control parametrisation [20], which is used to approximate the optimal control problem by a constrained non-linear programming problem. This software is backed by several theoretical advancements over earlier versions, such as superior technique for handling continuous state inequality constraints [19] and its ability to solve Koh's study ([9], [10], [11]) on optimising performance for the Yurchenko layout vault, a complex optimal control problem. In addition, it includes software hooks to four optimization algorithms namely, FSQP, NLPQL, NPSOL and NLPQLP. All four algorithms use a sequential quadratic programming algorithm which is recognised as the most efficient algorithm for small and medium size non-linearly constrained optimization problem [8]. MATLAB based systems for example, [5], which became available after the start of this project were considered too slow even though the literature suggests good performance.

The rest of this paper is organised as follows. In the next section, we introduce the geometry used for the paper. In Section 3, we propose new constraints, different from what was explored in Tan et al. [18], introduce the CPET technique and velocity decision variable within the objective function. The introduction of the mention will be used to sought better periodicity and different walking speeds. Numerical experimental results will be presented to demonstrate that this modified model, as compared to the model seen in Tan et al. [18], improves periodic motion and allow analysis of different walking speeds, and hence provide a more realistic tool for human walking modelling.

## 2 The Model

### 2.1 Geometry

A link segment model was used to represent the human body as depicted in Figure 2.1 in which there are $n = 7$ segments. For each $i = 1, \ldots, n$, the $i^{th}$ segment has length $\ell_i$, mass $m_i$ and moment of inertia $I_i$ about its center of mass (CoM)(within the segment), which is a distance $r_i$ from proximal $(x_i^p,\, y_i^p)$ and distance $l_i - r_i$ from the distal $(x_i^d,\, y_i^d)$ end (Figure 2.2). The link segment model is that used in Tan et al.'s [18].

Each segment's CoM position is defined by $(x_i,\, y_i)$, and angle $\theta_i$, where $\theta_i$ is determined by angle that segment makes with the positive $x$-axis. The CoM of the whole body, $(X,\, Y)$,

DISTRIBUTION A. Approved for public release: distribution unlimited.

is given by

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \frac{1}{M} \begin{bmatrix} \sum_{i=1}^{n} m_i x_i \\ \sum_{i=1}^{n} m_i y_i \end{bmatrix} \text{ where } M = \sum_{i=1}^{n} m_i.$$

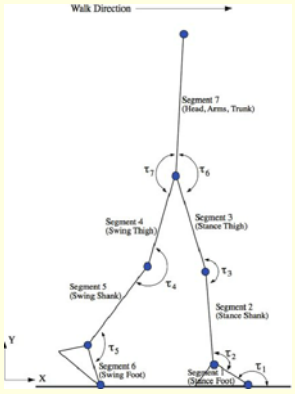



Figure 2.1: *Seven-segment model*

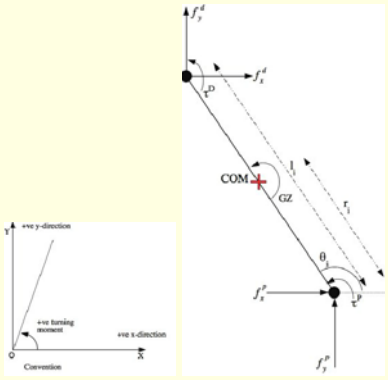



Figure 2.2: *An $i^{th}$ segment diagram*

The positional equations for a chain of segments (see Figure 2.1) are

$$\boldsymbol{x} = x_1^p \boldsymbol{e} + \boldsymbol{L}\boldsymbol{D}_c\boldsymbol{e}, \qquad \boldsymbol{y} = y_1^p \boldsymbol{e} + \boldsymbol{L}\boldsymbol{D}_s\boldsymbol{e},$$

where $\boldsymbol{x} = (x_1, \ldots, x_n)^t$, $\boldsymbol{y} = (y_1, \ldots, y_n)^t$, $\boldsymbol{e} = (1, 1, \ldots, 1)^t$,

$$\boldsymbol{D}_c = \operatorname{diag}(\cos\theta_1, \ldots, \cos\theta_n), \quad \boldsymbol{D}_s = \operatorname{diag}(\sin\theta_1, \ldots, \sin\theta_n)$$

and

$$\boldsymbol{L} = \begin{bmatrix} r_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ l_1 & r_2 & 0 & 0 & 0 & 0 & 0 \\ l_1 & l_2 & r_3 & 0 & 0 & 0 & 0 \\ l_1 & l_2 & l_3 & r_4 & 0 & 0 & 0 \\ l_1 & l_2 & l_3 & l_4 & r_5 & 0 & 0 \\ l_1 & l_2 & l_3 & l_4 & l_5 & r_6 & 0 \\ l_1 & l_2 & l_3 & 0 & 0 & 0 & r_7 \end{bmatrix}.$$

The CoM of the whole system is written in matrix-vector form as $MX = \boldsymbol{m}^t\boldsymbol{x}$ and $MY = \boldsymbol{m}^t\boldsymbol{y}$, where $\boldsymbol{m}^t = (m_1, \ldots, m_n)$. Hence the relations, using $\boldsymbol{m}^t\boldsymbol{e} = M$,

$$MX = \boldsymbol{m}^t\boldsymbol{x} = Mx_1^p + \boldsymbol{m}^t\boldsymbol{L}\boldsymbol{D}_c\boldsymbol{e} \quad \text{and} \quad MY = \boldsymbol{m}^t\boldsymbol{y} = My_1^p + \boldsymbol{m}^t\boldsymbol{L}\boldsymbol{D}_s\boldsymbol{e}.$$

The distal end of Segment 6 of the chain of segments, has co-ordinates

$$x_6^d = x_1^p + \sum_{i=1}^{6} l_i \cos\theta_i \quad \text{and} \quad y_6^d = y_1^p + \sum_{i=1}^{6} l_i \sin\theta_i.$$

DISTRIBUTION A. Approved for public release: distribution unlimited.

## 2.2 The Topology

The proximal incidence matrix is a $j \times n$ matrix $\boldsymbol{A}^p$ where

$$A^p_{ki} = \begin{cases} -1, & \text{if segment } i \text{ has proximal end at joint } k, \\ 0, & \text{otherwise.} \end{cases}$$

The distal incidence matrix $A^d$ is similarly defined,

$$A^d_{ki} = \begin{cases} 1, & \text{if segment } i \text{ has distal end at joint } k, \\ 0, & \text{otherwise.} \end{cases}$$

The joint-external contact incidence matrix $\boldsymbol{B}$ $(j \times e)$ is defined as

$$B_{ki} = \begin{cases} 1, & \text{if joint } k \text{ contacts the ground at external contact } i, \\ 0, & \text{otherwise,} \end{cases}$$

where $e$ is the number of external contacts.

Forces which supply the rotational and translational motions to segment $i$ and are given by

$$\boldsymbol{f}^e_k = \begin{bmatrix} f^{ex}_k \\ f^{ey}_k \end{bmatrix}, \quad \boldsymbol{f}^p_i = \begin{bmatrix} f^{px}_i \\ f^{py}_i \end{bmatrix}, \quad \boldsymbol{f}^d_i = \begin{bmatrix} f^{dx}_i \\ f^{dy}_i \end{bmatrix}.$$

The translational equations can be written as

$$-\boldsymbol{m}\dot{u}_1 + \boldsymbol{J}^y\dot{\boldsymbol{\omega}} + \boldsymbol{S}\boldsymbol{f}^x = -\boldsymbol{J}^x\boldsymbol{\omega}^2,$$
$$-\boldsymbol{m}\dot{v}_1 - \boldsymbol{J}^x\dot{\boldsymbol{\omega}} + \boldsymbol{S}\boldsymbol{f}^y = g\boldsymbol{m} - \boldsymbol{J}^y\boldsymbol{\omega}^2,$$

where proximal, distal and external forces are ordered such that

$$\boldsymbol{f}^x = \begin{bmatrix} \boldsymbol{f}^{px} \\ -\boldsymbol{f}^{ex} \\ \boldsymbol{f}^{dx} \end{bmatrix}, \qquad \boldsymbol{f}^y = \begin{bmatrix} \boldsymbol{f}^{py} \\ -\boldsymbol{f}^{ey} \\ \boldsymbol{f}^{dy} \end{bmatrix}.$$

We have defined: $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^t$, $\boldsymbol{\omega} = \dot{\boldsymbol{\theta}}$, $\dot{\boldsymbol{\omega}} = \ddot{\boldsymbol{\theta}}$, $\boldsymbol{\omega}^2 = (\omega_1^2, \ldots, \omega_n^2)^t$, $\boldsymbol{S} = [\boldsymbol{I}_n, \boldsymbol{0}, \boldsymbol{I}_n]$.

The moment equation can similarly be expressed in matrix form, with vector $\boldsymbol{\tau}$ being a vector of the proximal torques appropriately ordered, as

$$\boldsymbol{J}\dot{\boldsymbol{\omega}} + \boldsymbol{M}^x\boldsymbol{f}^x + \boldsymbol{M}^y\boldsymbol{f}^y = \boldsymbol{T}\boldsymbol{\tau},$$

where $\boldsymbol{J} = \operatorname{diag}(I_1, I_2, \ldots, I_n)$,

$$\boldsymbol{M}^x = \boldsymbol{D}_s \begin{bmatrix} -\boldsymbol{D}_r & | & \boldsymbol{0} & | & \boldsymbol{D}_l - \boldsymbol{D}_r \end{bmatrix},$$
$$\boldsymbol{M}^y = \boldsymbol{D}_c \begin{bmatrix} \boldsymbol{D}_r & | & \boldsymbol{0} & | & -(\boldsymbol{D}_l - \boldsymbol{D}_r) \end{bmatrix},$$
$$\boldsymbol{D}_r = \operatorname{diag}(r_1, r_2, \ldots, r_n), \ \boldsymbol{D}_l = \operatorname{diag}(l_1, l_2, \ldots, l_n).$$

The matrix $\boldsymbol{T}$ with a torque acting between Segment 6 and the external world is given by

$$\boldsymbol{T} = \begin{array}{c} \\ \text{seg1} \\ \text{seg2} \\ \text{seg3} \\ \text{seg4} \\ \text{seg5} \\ \text{seg6} \\ \text{seg7} \end{array} \begin{array}{c} \begin{array}{ccccccc} \tau_1 & \tau_2 & \tau_3 & \tau_4 & \tau_5 & \tau_6 & \tau_7 \end{array} \\ \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix} \end{array}.$$

DISTRIBUTION A. Approved for public release: distribution unlimited.

The complete equations for non-heel contact are:

$$\dot{\boldsymbol{\theta}} = \boldsymbol{\omega},$$

$$\begin{bmatrix} \dot{x}_1^p \\ \dot{y}_1^p \end{bmatrix} = \begin{bmatrix} u_1 \\ v_1 \end{bmatrix},$$

$$\begin{bmatrix} \dot{x}_6^d \\ \dot{y}_6^d \end{bmatrix} = \begin{bmatrix} u_6 \\ v_6 \end{bmatrix},$$

$$\begin{array}{cc} & \begin{array}{ccccccccc} n & 1 & 1 & 1 & 1 & n & n-1 & n & n-1 \end{array} \\ \begin{array}{c} n \\ 1 \\ 1 \\ (a)\ 1 \\ (a)\ 1 \\ (b)\ 1 \\ (b)\ 1 \\ n \\ n-2 \\ n \\ n-2 \end{array} & \begin{bmatrix} \boldsymbol{J} & 0 & 0 & 0 & 0 & \boldsymbol{M}^{px} & \boldsymbol{M}^{dx} & \boldsymbol{M}^{py} & \boldsymbol{M}^{dy} \\ \boldsymbol{l}_s^t & -1 & 0 & 1 & 0 & \boldsymbol{0}^t & \boldsymbol{0}^t & \boldsymbol{0}^t & \boldsymbol{0}^t \\ -\boldsymbol{l}_c^t & 0 & -1 & 0 & 1 & \boldsymbol{0}^t & \boldsymbol{0}^t & \boldsymbol{0}^t & \boldsymbol{0}^t \\ [\boldsymbol{0}^t & 1 & 0 & 0 & 0 & \boldsymbol{0}^t & \boldsymbol{0}^t & \boldsymbol{0}^t & \boldsymbol{0}^t] \\ [\boldsymbol{0}^t & 0 & 1 & 0 & 0 & \boldsymbol{0}^t & \boldsymbol{0}^t & \boldsymbol{0}^t & \boldsymbol{0}^t] \\ [\boldsymbol{0}^t & 0 & 0 & 1 & 0 & \boldsymbol{0}^t & \boldsymbol{0}^t & \boldsymbol{0}^t & \boldsymbol{0}^t] \\ [\boldsymbol{0}^t & 0 & 0 & 0 & 1 & \boldsymbol{0}^t & \boldsymbol{0}^t & \boldsymbol{0}^t & \boldsymbol{0}^t] \\ \boldsymbol{J}^y & -\boldsymbol{m} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & -\boldsymbol{A}^p & \boldsymbol{A}^d & \boldsymbol{0} & \boldsymbol{0} \\ -\boldsymbol{J}^x & \boldsymbol{0} & -\boldsymbol{m} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{I} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & -\boldsymbol{A}^p & \boldsymbol{A}^d \end{bmatrix} \end{array} \begin{bmatrix} \dot{\boldsymbol{\omega}} \\ \dot{u}_1 \\ \dot{v}_1 \\ \dot{u}_6 \\ \dot{v}_6 \\ \boldsymbol{f}^{px} \\ \boldsymbol{f}^{dx} \\ \boldsymbol{f}^{py} \\ \boldsymbol{f}^{dy} \end{bmatrix}$$

$$= \begin{bmatrix} \boldsymbol{T}\boldsymbol{\tau} \\ -\boldsymbol{l}_c^t\boldsymbol{\omega}^2 \\ -\boldsymbol{l}_s^t\boldsymbol{\omega}^2 \\ 0 \\ 0 \\ 0 \\ 0 \\ -\boldsymbol{J}^x\boldsymbol{\omega}^2 \\ \boldsymbol{0} \\ -\boldsymbol{J}^y\boldsymbol{\omega}^2 + \boldsymbol{m}g \\ \boldsymbol{0} \end{bmatrix}.$$

In above equation we used double square brackets to differentiate these rows from the others. Depending on the cases determined by the phase the walking motion is in, some of these rows are not needed. For more details on the cases we refer to [18]. The two row vectors $\boldsymbol{l}_s^t$ and $\boldsymbol{l}_c^t$ in the above equation are defined as

$$ {}_1\boldsymbol{l}^t = [l_1, l_2, l_3, l_4, l_5, l_6, 0], \quad \text{or} \quad {}_2\boldsymbol{l}^t = [0, l_2, l_3, l_4, l_5, l_6, 0],$$

$$ {}_1\boldsymbol{l}_s^t = {}_1\boldsymbol{l}^t\boldsymbol{D}_s, \quad \text{or} \quad {}_2\boldsymbol{l}_s^t = {}_2\boldsymbol{l}^t\boldsymbol{D}_s,$$

$$ {}_1\boldsymbol{l}_c^t = {}_1\boldsymbol{l}^t\boldsymbol{D}_c, \quad \text{or} \quad {}_2\boldsymbol{l}_c^t = {}_2\boldsymbol{l}^t\boldsymbol{D}_c,$$

where for proximal Segment 1

$$\begin{bmatrix} x_6^d \\ y_6^d \end{bmatrix} = \begin{bmatrix} x_1^p \\ y_1^p \end{bmatrix} + \begin{bmatrix} {}_1\boldsymbol{l}_c^t\boldsymbol{e} \\ {}_1\boldsymbol{l}_s^t\boldsymbol{e} \end{bmatrix}, \qquad \begin{bmatrix} \dot{x}_6^d \\ \dot{y}_6^d \end{bmatrix} = \begin{bmatrix} \dot{x}_1^p \\ \dot{y}_1^p \end{bmatrix} + \begin{bmatrix} -{}_1\boldsymbol{l}_s^t\boldsymbol{\omega} \\ {}_1\boldsymbol{l}_c^t\boldsymbol{\omega} \end{bmatrix},$$

DISTRIBUTION A. Approved for public release: distribution unlimited.

and renaming the velocities as $u$ and $v$, scripted appropriately,

$$\begin{bmatrix} \dot{u}_6^d \\ \dot{v}_6^d \end{bmatrix} = \begin{bmatrix} \dot{u}_1^p \\ \dot{v}_1^p \end{bmatrix} + \begin{bmatrix} -{}_1\boldsymbol{l}_s^t\dot{\boldsymbol{\omega}} \\ {}_1\boldsymbol{l}_c^t\dot{\boldsymbol{\omega}} \end{bmatrix} + \begin{bmatrix} -{}_1\boldsymbol{l}_c^t\boldsymbol{\omega}^2 \\ -{}_1\boldsymbol{l}_s^t\boldsymbol{\omega}^2 \end{bmatrix}.$$

Similarly for distal Segment 6 measured from proximal Segment 2.

## 3 Numerical Techniques

The simulation model (Experiment Main), that was explored in Tan et al.'s [18] paper, was extended to conduct three other experiments. These experiments were based on optimized toques obtained from Experiment Main. The initial study presents 18 states ($\boldsymbol{x} = [x_1, x_2, \ldots, x_{18}]^\top$), 15 system parameters ($\boldsymbol{z} = [z_1, z_2, \ldots, z_{15}]^\top$), and 7 controls ($\boldsymbol{\tau} = [\tau_1, \tau_2, \ldots, \tau_7]^\top$) namely joint torques, were set up in MISER3.3 [8]. The 18 states consists of the angular displacements from Segment 1 to Segment 7 ($x_i = \theta_i, i = 1, \ldots, 7$), angular velocity ($x_i = \theta_i, i = 8, \ldots, 14$), coordinate and velocity of proximal end of segment one, $((x_{15}, x_{16}, x_{17}, x_{18}) = (x_1^p, y_1^p, \dot{x}_1^p, \dot{y}_1^p))$. The system parameter consists of the initial segment angular orientation ($\theta_i(0) = z_i, i = 1, \ldots, 7$), initial angular velocity at start of single support phase ($\omega_i(0) = z_{7+i}$) and $z_{15}$ is the step length which is twice the distance of initial distance of $x_1^p(0)$ and $x_6^d(0)$ and hence dependent on $(z_1, \ldots, z_6)$. Variables $\theta_1, \theta_2, \theta_3$, $\theta_4, \theta_5, \theta_6$ describe the angular displacements of the legs and $\theta_7$ describes the angular displacement of the trunk segment. $\omega_1, \ldots, \omega_7$ are the segments' corresponding velocities; and $(x_1^p, y_1^p)$ are the coordinates of the proximal end of segment one (toe of stance foot) which remains stationary on the ground during one step of the walk cycle.

Normal walking has been assumed to be symmetric and cyclic and hence only one step of the gait cycle needs to be modelled and simulated. Periodicity conditions are required such that the end of the walk cycle is identical to the start so that successive steps repeat the motion of the previous step by swapping the roles of legs.

Experiment 1 is an extension from the main experiment, aimed at achieving a periodic motion. The objective function remained the same as the main experiment,

$$G_0(\boldsymbol{\tau}, \boldsymbol{z}) \quad = \quad \int_0^{T_f} (CoM_{ypos} - CoM_{yinit})^2 \, dt$$

with the following state equations:

$$\dot{\boldsymbol{x}}(t) = \begin{cases} \boldsymbol{f}_1(t, \boldsymbol{x}, \boldsymbol{u}, \boldsymbol{z}), & t \in [0, T_1), \boldsymbol{x}(0) = \boldsymbol{x}^0(\boldsymbol{z}), \\ \boldsymbol{f}_2(t, \boldsymbol{x}, \boldsymbol{u}, \boldsymbol{z}), & t \in [T_1, T_f), \boldsymbol{x}(T_1) = \boldsymbol{h}_1(\boldsymbol{x}(T_1^-), \boldsymbol{z}), \end{cases}$$

where $T_1$ $(= 0.386s)$ is the duration of the single support phase, $T_f$ $(= 0.486s)$ is the duration of a step (single support and double support phase) and $\boldsymbol{h}_1(\boldsymbol{x}(T_1^-), \boldsymbol{z})$ defines the new states governing the start of double support phase. $CoM_{ypos}$ is the center of mass of y-coordinate, a function of $\boldsymbol{\theta}(\boldsymbol{\tau}, \boldsymbol{z}, t)$, and $CoM_{yinit}$ is the initial center of mass of y-coordinate, a function of $\boldsymbol{z}$, as calculated by MISER3.3.

The objective function is subject to constraints in the canonical form:

$$G_k(\boldsymbol{u}, \boldsymbol{z}) = \phi_k(\boldsymbol{x}(t_k, \boldsymbol{z})) + \int_0^{t_k} g_k(t, \boldsymbol{x}(t), \boldsymbol{u}(t), \boldsymbol{z}) \, \mathrm{d}t \quad \left.\begin{matrix} = \\ \geq \end{matrix}\right\} 0, k = 1, \ldots, n_{gc},$$

where $n_{gc}$ is the total number of canonical constraints, and $t_k \in (0, t_f]$ is a known constant and is referred to as the 'characteristic time' associated with the constraint $G_k$. All-time

DISTRIBUTION A. Approved for public release: distribution unlimited.

constraints $h(t, \boldsymbol{x}, \boldsymbol{u}, \boldsymbol{z}) \geq 0$ and constraints involving system parameters $g_k(\boldsymbol{z})$ as well, are converted by MISER3.3 to canonical constraints. The gradient of the objective and constraint functions are automatically calculated by MISER3.3 using a numerical procedure. See [8] for more details.

Experiments 1 follow the same constraints as Experiment Main (See [18] for more details) with the exception of the terminal constraint on the trunk angular displacement at ($T_f = 0.486s$)

$$g_4 \equiv 0, \qquad \phi_4(\boldsymbol{\theta}(T_f), \boldsymbol{z}) = z_7 - \theta_7(T_f) = 0.$$

which is now replaced with,

- a terminal constraint at the end of the step cycle ($T_f = 0.486s$) on the final $\boldsymbol{\theta}(T_f)$, such that the final motion resembles the first, is given by,

$$g_4 \equiv 0, \qquad \phi_4(\boldsymbol{\theta}(T_f), \boldsymbol{z}) = \sum_{i=1}^{3}(z_i - \theta_{7-i}(T_f) + 3.14)^2 + \sum_{i=4}^{5}(z_i - \theta_{7-i}(T_f) - 3.14)^2 + (z_7 - \theta_7(T_f))^2 = 0.$$

Remark: Once again, a sum of squares of 7 constraints is taken to reduce complexity.

Experiments 2 and 3 aimed at investigating at maximising and minimising velocity of normal walking and observing the motion and joint torques involved respectively. This was done using the Control Parametrisation Enhancing Transform (CPET) technique for constrained optimal control problems in MISER3 [21] that allowed the time of a step cycle to be optimized and adjusting the time interval of single support and double support accordingly. CPET is a technique which can be used to optimise the real time taken for a set of states to move from one configuration to another by introducing a new control function $u_8(t) = \frac{ds}{dt}$, modelled as a piecewise constant control function on user chosen knots, in this case $(\xi_0, \xi_1, \xi_3) = (0, T_1, T_f)$. In addition, an extra state function $x_{19}$ is added to the state variables with different equation,

$$\frac{ds(t)}{dt} = u_8(t), \qquad x_{19}(0) = z_{16}.$$

The optimal control problem is redefined with 19 states ($\hat{\boldsymbol{x}} = [\boldsymbol{x}(x_{19}), x_{19}]$), 16 system parameters ($\hat{\boldsymbol{z}} = [\boldsymbol{z}, z_{16}]$), and 7 controls ($\hat{\boldsymbol{u}} = [\boldsymbol{u}(x_{19}), u_8]$). For experiment 2 and 3, the revised objective function is,

$$G_0(\hat{\boldsymbol{\tau}}, \boldsymbol{z}) = \int_0^1 u_8(t)(CoM_{ypos} - CoM_{yinit})^2\, dt$$

with refined dynamics

$$\frac{d\boldsymbol{x}}{dt} = \begin{cases} \begin{pmatrix} u_8(t)\boldsymbol{f}_1(x_{15}(t), \hat{\boldsymbol{x}}, \hat{\boldsymbol{u}}, \boldsymbol{z}) \\ u_8(t) \end{pmatrix}, & t \in [0, t_k), \hat{\boldsymbol{x}}(0) = \begin{pmatrix} \boldsymbol{x}^0(\boldsymbol{z}) \\ z_{16} \end{pmatrix}, \\ \begin{pmatrix} u_8(t)\boldsymbol{f}_2(x_{15}(t), \boldsymbol{x}, \boldsymbol{u}, \boldsymbol{z}) \\ u_8(t) \end{pmatrix}, & t \in [t_k, 1), \hat{\boldsymbol{x}}(t_k) = \begin{pmatrix} \boldsymbol{h}_1(\boldsymbol{x}(t_k^-), \boldsymbol{z}) \\ x_{15}(t_k^-) \end{pmatrix}, \end{cases}$$

DISTRIBUTION A. Approved for public release: distribution unlimited.

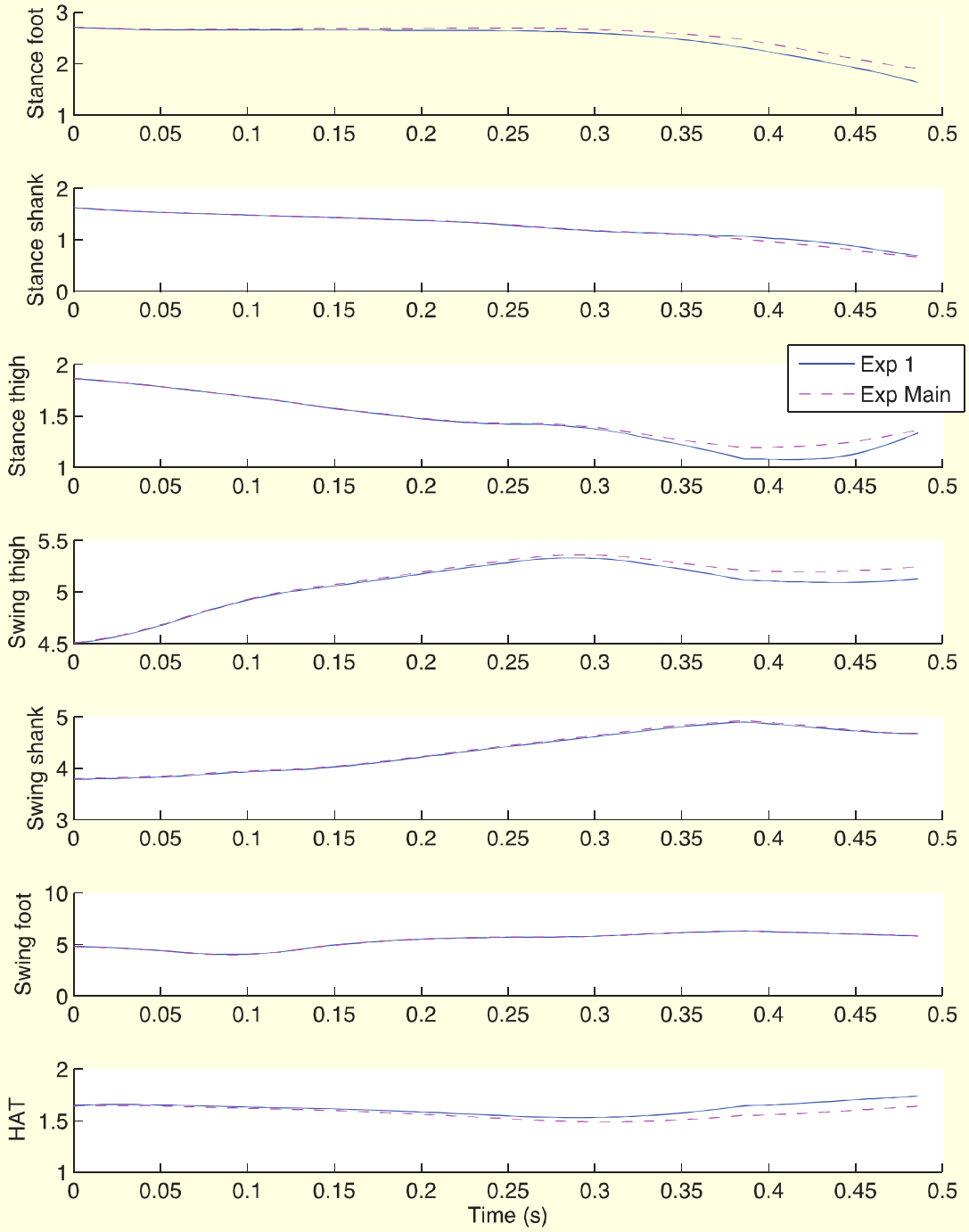



Figure 4.3: *Comparison of segment angular displacements between Exp1 and Exp Main*

Figure 4.5 presents the external forces acting on the ankle during double support phase. A negative external vertical force is observed when swing heel comes into contact with the

**DISTRIBUTION A. Approved for public release: distribution unlimited.**

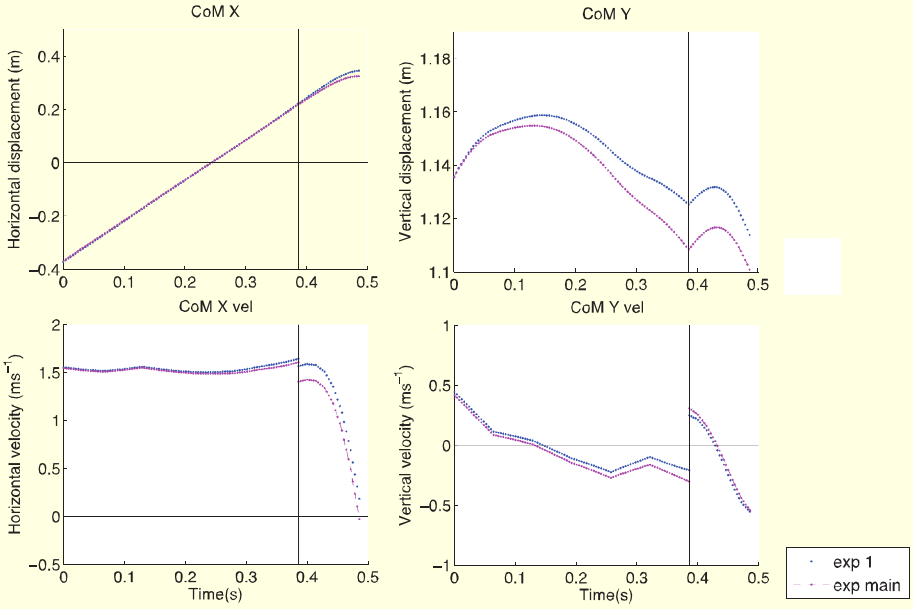



Figure 4.4: *Exp 1 versus Exp Main - CoM displacement and velocity*

ground, which is consistent as weight of the body is being distributed from the stance toe at start and through to double support phase. During this period, a positive horizontal external force keeps the heel in position, preventing it from sliding backwards. When the present model is compared, it can be observed that there is none or only slight difference in the horizontal external force on the ankle. However, in contrast, the present model presents a larger negative vertical force as compared to Experiment Main. A closer look at the proximal and distal $y$-forces during double support phase suggest that greater force in the $y$-direction is present in this model (Figures 4.6 and 4.7), and is especially evident in the forces acting on the swing leg.

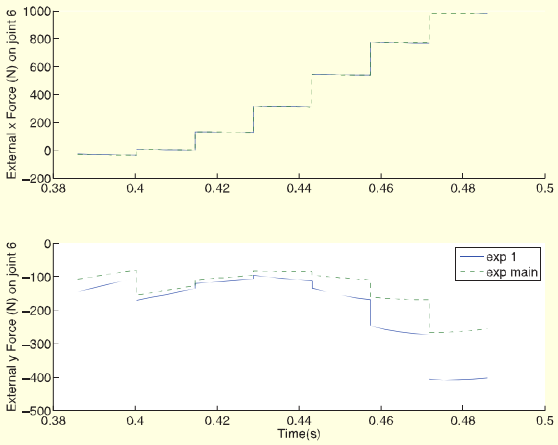



Figure 4.5: *Exp 1 versus Exp Main - External forces (x, y) from ankle to ground during double support phase*

DISTRIBUTION A. Approved for public release: distribution unlimited.

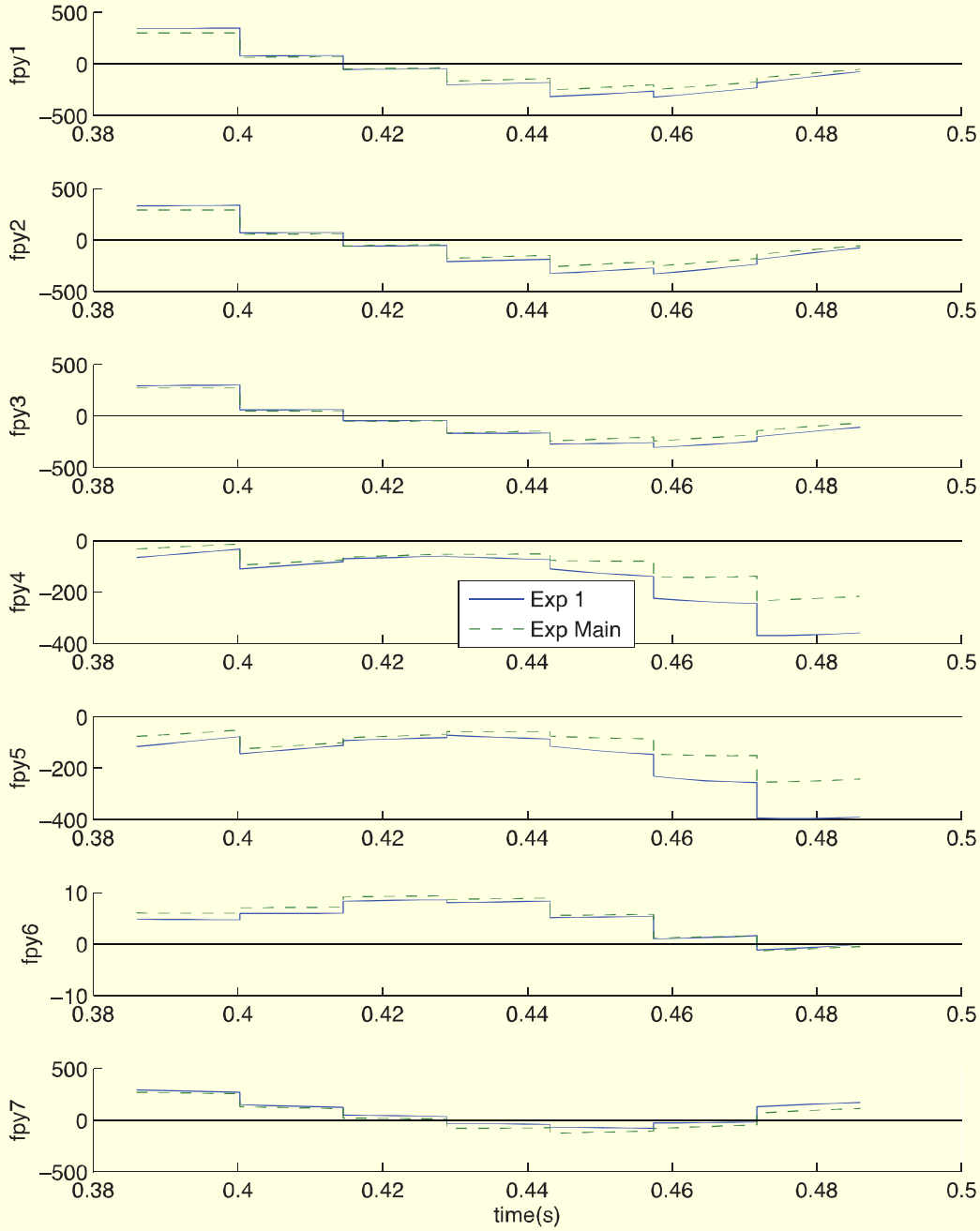



Figure 4.6: *Exp 1 versus Exp Main - Proximal y forces (N) during double support phase*

DISTRIBUTION A. Approved for public release: distribution unlimited.

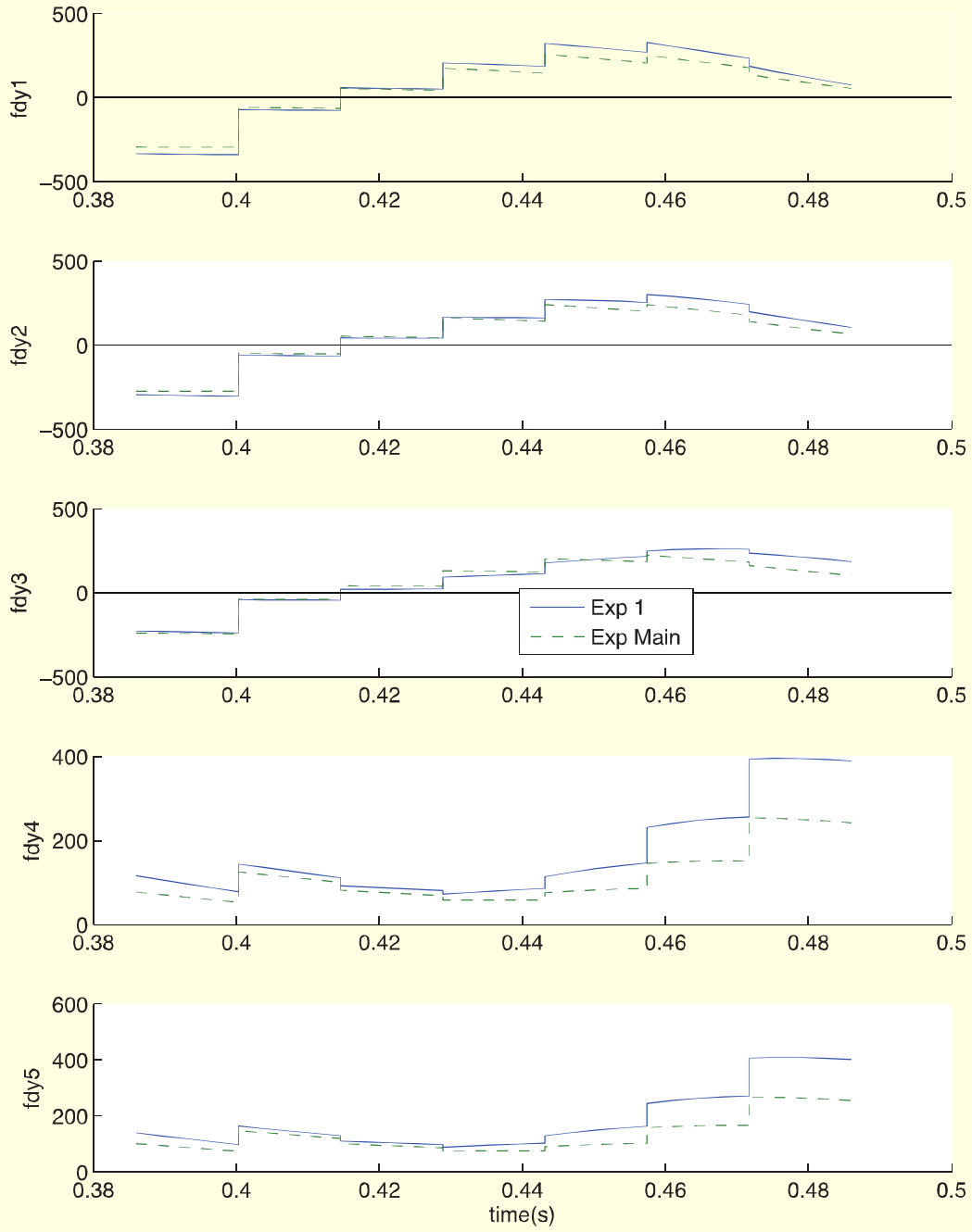



Figure 4.7: *Exp 1 versus Exp Main - -Distal y forces (N) during double support phase*

DISTRIBUTION A. Approved for public release: distribution unlimited.

Periodicity constraint is a terminal time constraint which is only active at the end of the time interval ($t = 0.486s$) in Experiment 1. It was observed that the addition of periodicity constraint does improve periodic motion in the model, with adjustments in segment angular displacements occurring mainly during the double support phase. In addition, it improved the objective, keeping the $y$-CoM closer to initial. Very small changes or even sometimes none were found in the torques and horizontal forces between Experiment Main and Experiment 1. The main components which influenced periodicity were observed to be the $y$-component forces acting on the swing leg during double support phase, although no changes or only slight changes were observed during the single support phase. This is not surprising since this constraint is only active at final time, and adjustments are only required closer to $T_f = 0.486s$. A larger vertical force observed is possibly due to keeping the body upright, and $y$-CoM throughout closer to its initial position.

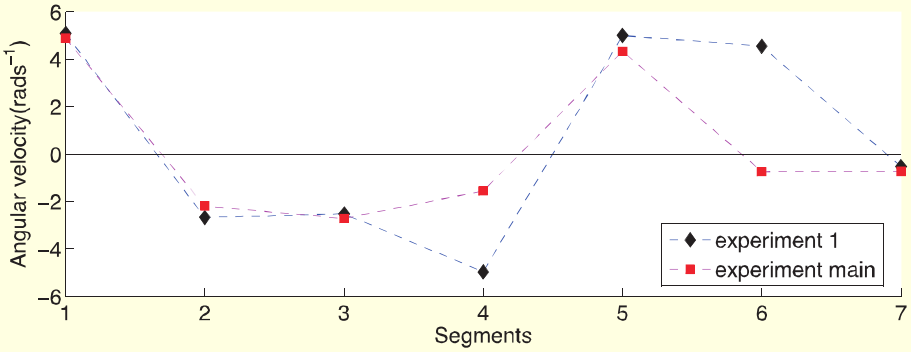



Figure 4.8: *Exp 1 versus Exp Main - Comparison of velocity periodicity differences*

A periodic motion was achieved by improving angular periodicity with the implementation of its constraints, however it comes with its flaws. Illustrated in Figure 4.8, larger velocity periodicity difference values was observed for Experiment 1 despite having angular periodic constraints in place. Having periodicity constraints to improve periodic motion may not be ideal afterall.

## 4.2 Experiment 2

Experiment 2 investigated if a maximum velocity can be achieved by extending Experiment Main. This was done by minimising time and maximising distance in the objective function using CPET (Section 2). The minimum time of a full walk cycle was calculated to be $t = 0.386s$, which is $0.1s$ shorter than the original time taken from the data. Figure 4.9 depicts the horizontal velocity of the CoM, where a faster velocity can be observed from Experiment 2 as compared to Experiment Main. Velocity increases towards the end of single support phase and a sudden drop in velocity occurs as swing leg hits the ground during the very short double support phase.

The result of Experiment 2 suggests that at faster walking speed, the ratio of single support phase in a full walk cycle is nearly one, which indicates that double support phase becomes almost negligible or can be considered instantaneous. Figure 4.10 illustrated a full walk cycle of Experiment 2. It was observed that the whole swing foot comes into contact with the ground at the same time instead of following a heel strike then toe strike motion.

The optimized joint torque trajectories for Experiment 2 are presented on Figure 4.11. As these trajectories were originally derived from the initial joint moment estimates computed

DISTRIBUTION A. Approved for public release: distribution unlimited.

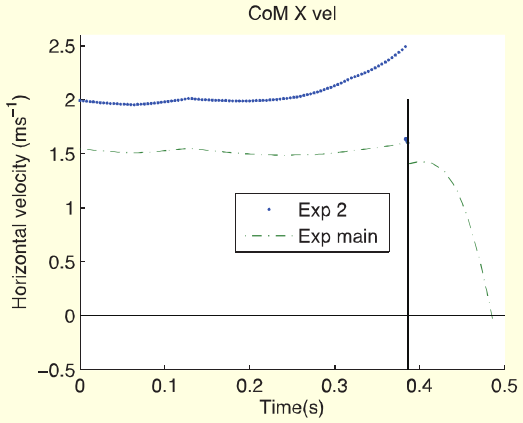



Figure 4.9: *Exp 2 versus Exp Main - Horizontal velocity of CoM*

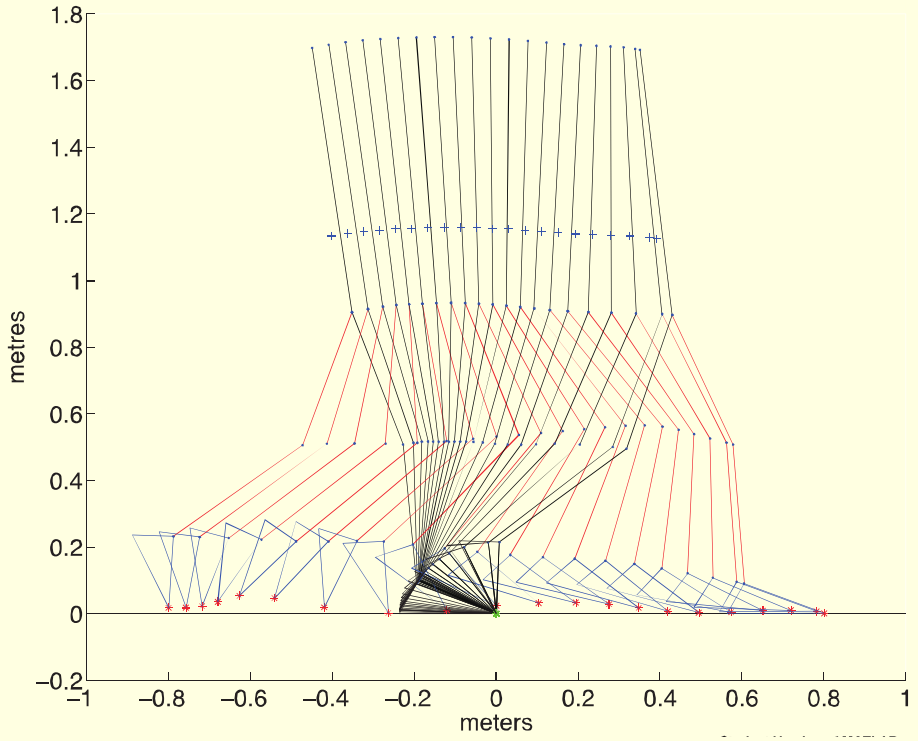



Figure 4.10: *One step walk motion - Experiment 2*

by the method of inverse dynamics, they are specific to the movement pattern. As compared to Experiment Main, in the single support phase, the torques behaved similarly. However, as the double support phase interval is much smaller, so are the knots interval for torques occurring during this period, hence torque changes rapidly so as to follow through the movement pattern at a quicker time. This set of torque trajectories is able to produce the same movement characteristics as in Experiment Main, depicted by the segment trajectories

DISTRIBUTION A. Approved for public release: distribution unlimited.

in Figure 4.12 but in a faster time frame. In addition, the horizontal distance of CoM in Experiment 2 travelled a slightly further distance than Experiment Main in a shorter time (Figure 4.13).

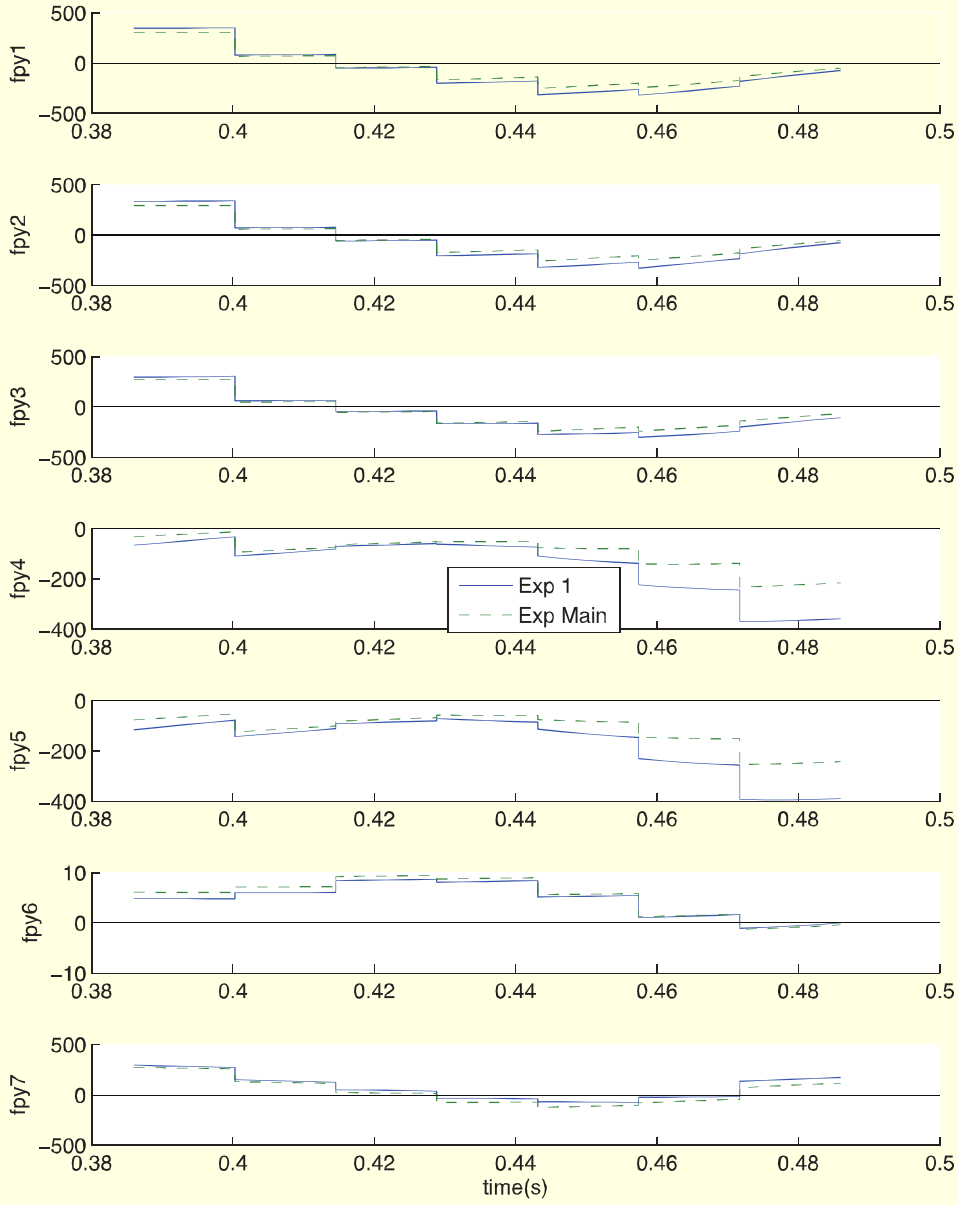



Figure 4.11: *Exp 2 versus Exp Main - optimized torques*

DISTRIBUTION A. Approved for public release: distribution unlimited.

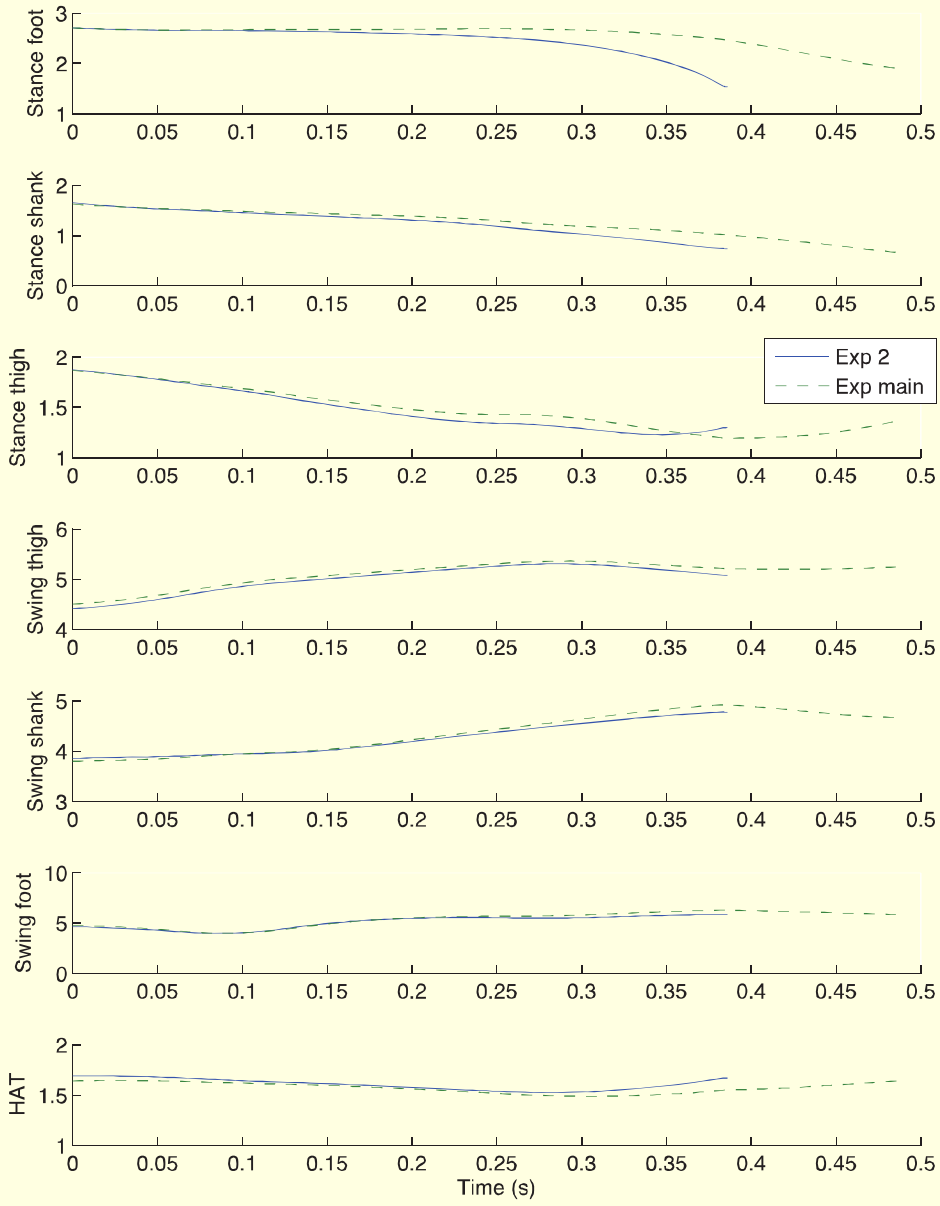



Figure 4.12: *Exp 2 versus Exp Main - Segment angular displacement trajectories*

DISTRIBUTION A. Approved for public release: distribution unlimited.

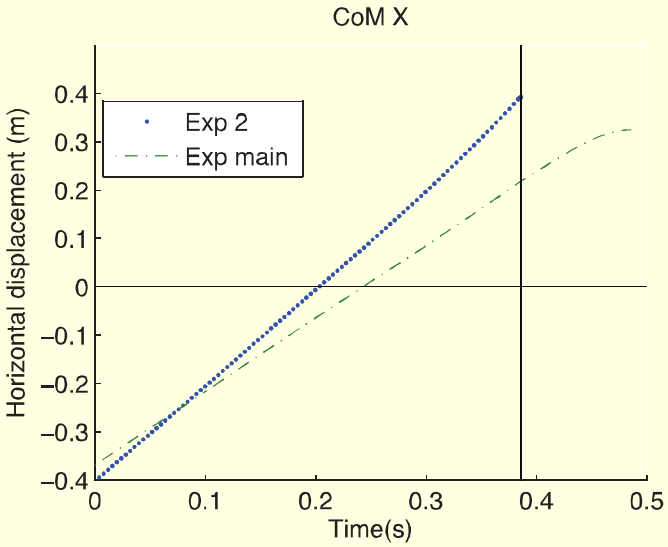



Figure 4.13: *Exp 2 versus Exp Main - Horizontal displacement of CoM*

Only trunk periodicity, following Experiment Main, was considered in Experiment 2. However, Experiment 2 achieved better angular periodicity than Experiment Main (Figure 4.14). Table 4.2 presents the periodicity difference values of Experiment Main, 1 and 2, where Experiment 1 considered periodicity constraints. The values indicated that periodicity constraints need not necessary be considered and periodic motion can be achieved in a fast walk. Figure 4.15 displays the start and end of the walk cycle, where a periodic motion can be observed.

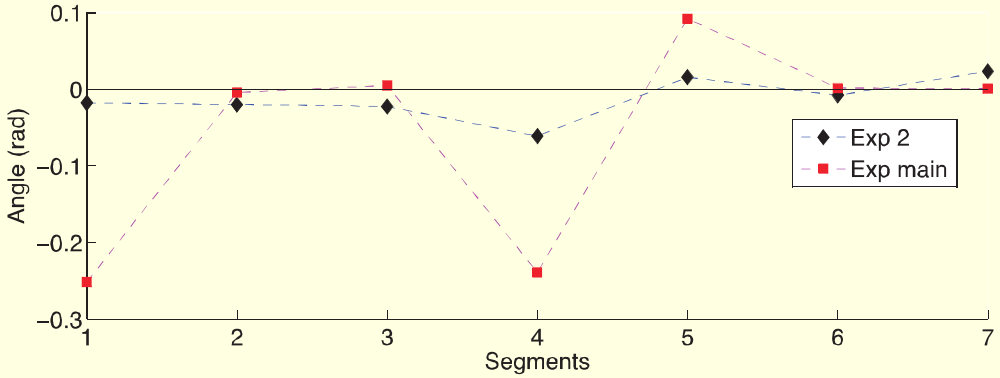



Figure 4.14: *Exp 2 versus Exp Main - Angular periodicity differences*

DISTRIBUTION A. Approved for public release: distribution unlimited.

| Periodicity | Experiment 2 | Experiment Main | Experiment 1 |
|---|---|---|---|
| $z_6 - \theta_1(t_f) - \pi$ | -0.0178 | -0.2513 | 0.0418 |
| $z_5 - \theta_2(t_f) - \pi$ | -0.0199 | -0.0043 | -0.0533 |
| $z_4 - \theta_3(t_f) - \pi$ | -0.0224 | 0.0048 | 0.0256 |
| $z_3 - \theta_4(t_f) + \pi$ | -0.0613 | -0.2388 | -0.1196 |
| $z_2 - \theta_5(t_f) + \pi$ | 0.0155 | 0.0920 | 0.0914 |
| $z_1 - \theta_6(t_f) + \pi$ | -0.0080 | 0.0017 | 0.0369 |
| $z_7 - \theta_7(t_f)$ | 0.0234 | 0.0006 | -0.0859 |

Table 4.2: *Angular periodicity differences*

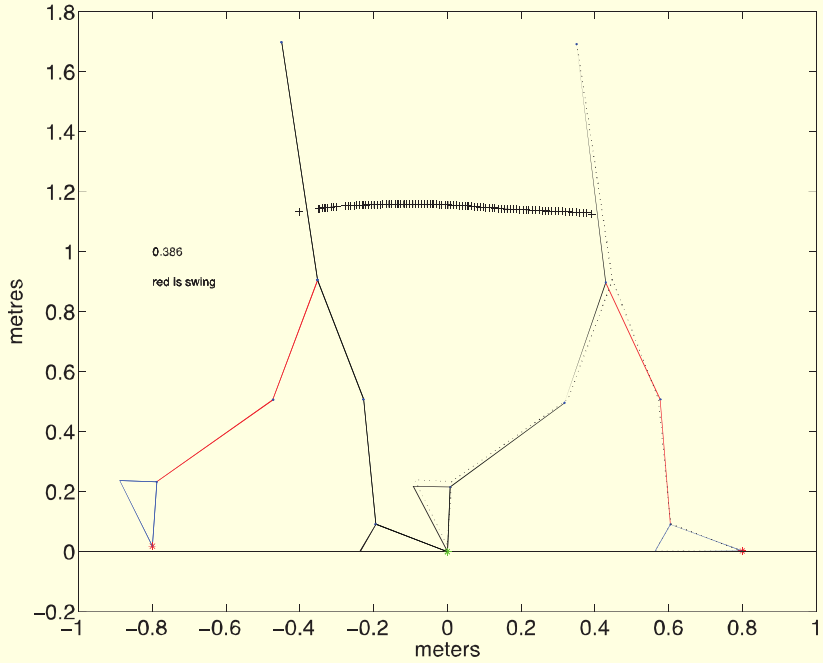



Figure 4.15: *Start and end of walk cycle with center of mass position - Experiment 2*

Table 4.2 provides a comparison of the optimized system parameters of Experiment 2 and Experiment Main initial angles and angular velocity. Differences of the initial angular displacements can be observed between Experiment 2 and Main. As maximising step length was part of the objective, system parameter $z_{15}$, defined to be half step length, had to be maximised. Adjusting $z_{15}$ to maximise step length, would change the initial distance between the stance toe and swing toe. As compared to Experiment Main, where $z_{15} = 0.7351m$, in order to satisfy objective, the maximum boundary of $z_{15}(0.8m)$ was reached in Experiment 2. Initial angular displacements had to be adjusted accordingly as $z_{15}$ changes in Experiment 2, in order for certain constraints to be satisfied such as, stance toe and swing toe had to be of distance $z_{15}$ while keeping the body upright.

The adjustments in initial angles could also be the explanation which resulted in periodicity being achieved as illustrated in Figure 4.15.

Vertical displacement of CoM for Experiment 2 was observed to be smaller than Experi-

DISTRIBUTION A. Approved for public release: distribution unlimited.

| Parameters | Experiment 2 | Experiment Main |
|---|---|---|
| $z_1$ | 2.7055 | 2.7055 |
| $z_2$ | 1.6499 | 1.6269 |
| $z_3$ | 1.8751 | 1.8656 |
| $z_4$ | 4.4185 | 4.5093 |
| $z_5$ | 3.8573 | 3.8011 |
| $z_6$ | 4.6592 | 4.7978 |
| $z_7$ | 1.6927 | 1.6411 |
| $z_8$ | -1.1615 | -1.2979 |
| $z_9$ | -3.1145 | -2.4389 |
| $z_{10}$ | -1.6123 | -1.3727 |
| $z_{11}$ | 1.7188 | 1.6809 |
| $z_{12}$ | 1.2754 | 1.1253 |
| $z_{13}$ | -4.0441 | -4.0952 |
| $z_{14}$ | 0.0571 | 0.3605 |
| $z_{15}$ | 0.8 | 0.7351 |

Table 4.3: *System parameters*

ment Main (Figure 4.16). The difference between the vertical displacement of CoM through the full walk cycle and the initial position of $y$-CoM is lesser especially at final time since a periodic motion was achieved.

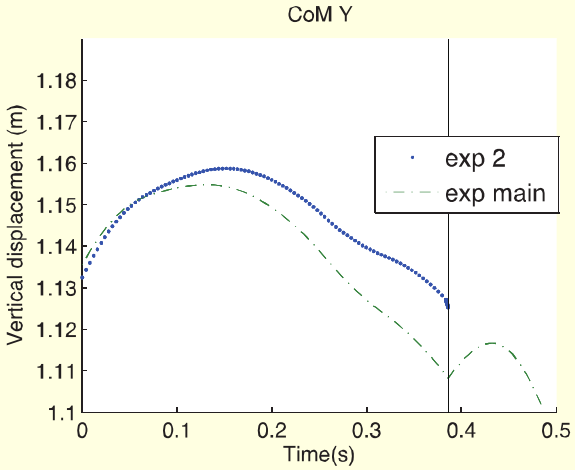



Figure 4.16: *Exp 2 versus Exp Main - Vertical displacement of CoM*

The results of Experiment 2 suggested that at faster walk speed, single support phase contributes to the majority of the walk cycle and double support phase may be deemed instantaneous. Forces and torques behaved similar to Experiment Main except during double support phase. During double support phase, forces and torques increase and decrease in the same direction, but due to the short time interval present in this phase, forces and torques increase and decrease rapidly. This rapid change allows the model to reach final position in a shorter time interval. The final result of Experiment 2 also suggested that periodic motion

DISTRIBUTION A. Approved for public release: distribution unlimited.

can be achieved at faster walking speed without the need of periodic constraints.

## 4.3 Experiment 3

Experiment 3 investigated if a minimum velocity can be achieved by extending Experiment Main. This was carried out by minimising distance and maximising time but with a lower bound on distance and upper bound on time. These limits were reached. The maximum time for a full walk cycle was restricted to be $t = 0.486s$ and step length to be $1.22m$, which is $0.25m$ shorter than the original step length. Figure 4.17 presents the horizontal velocity of CoM of Experiment 3. A slower velocity can be observed in comparison with Experiment Main for single support phase, but similar velocity is noted for double support phase. However, the jump in velocity was observed to occur earlier than $t = 0.386s$, which implies that the duration for single support phase is slightly shorter in Experiment 3 than Experiment Main and longer for double support phase. CPET (Section 2) is used to optimise the change over time. Time interval for single support phase in Experiment 3 ends earlier at $t = 0.3762$, which is $0.0098s$ shorter than Experiment Main, which lengthened double support phase.

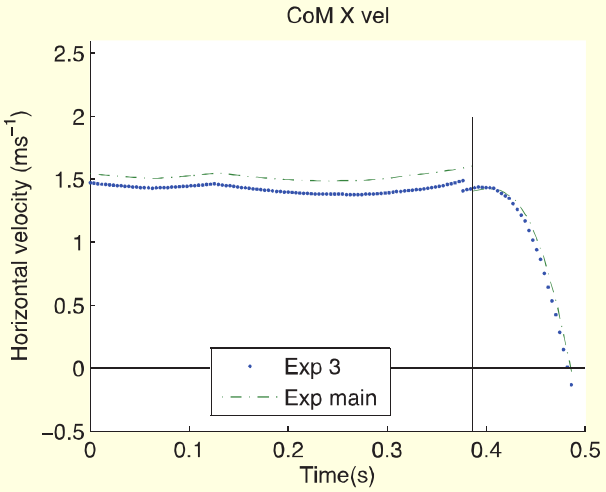



Figure 4.17: *Exp 3 versus Exp Main - Horizontal velocity of CoM*

The result suggest when minimising velocity, even though time to complete a full cycle remains the same $t = 0.486s$, the duration of each phases changed slightly with shorter step length distance. Since step length is smaller, swing heel strike occurred in a shorter time span, but duration of double support phase was lengthened in order to maximise time to complete the walk cycle. Figure 4.18 illustrates the horizontal displacement of CoM of Experiment 3 which was observed to be shorter as $t$ increases when compared to Experiment Main. Figure 4.19 depicts the walk motion of Experiment 3, where the walk is observed to be tighter and the distance between the swing foot and stance foot is smaller.

The optimized joint torque trajectories for Experiment 3 were observed to have similar pattern since it was derived from initial joint moment estimates computed by inverse dynamics. However, due to the change in time interval, shifts in joint torques towards the left was observed in single support phase, and towards the right in double support phase (Figure 4.20 and 4.21). During the single support phase, as time interval was smaller, knot intervals

DISTRIBUTION A. Approved for public release: distribution unlimited.

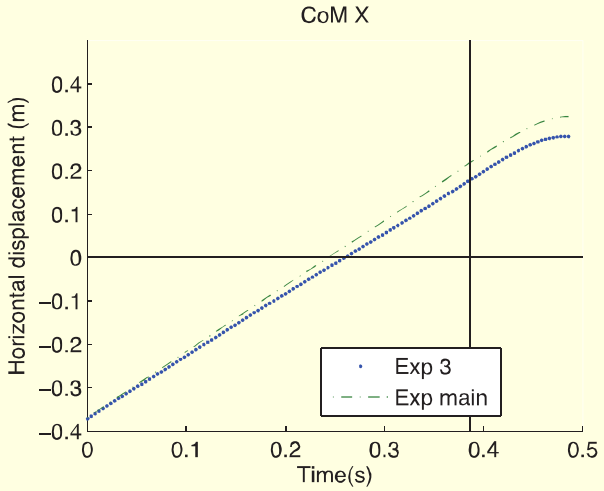



Figure 4.18: *Exp 3 versus Exp Main - Horizontal displacement of CoM*

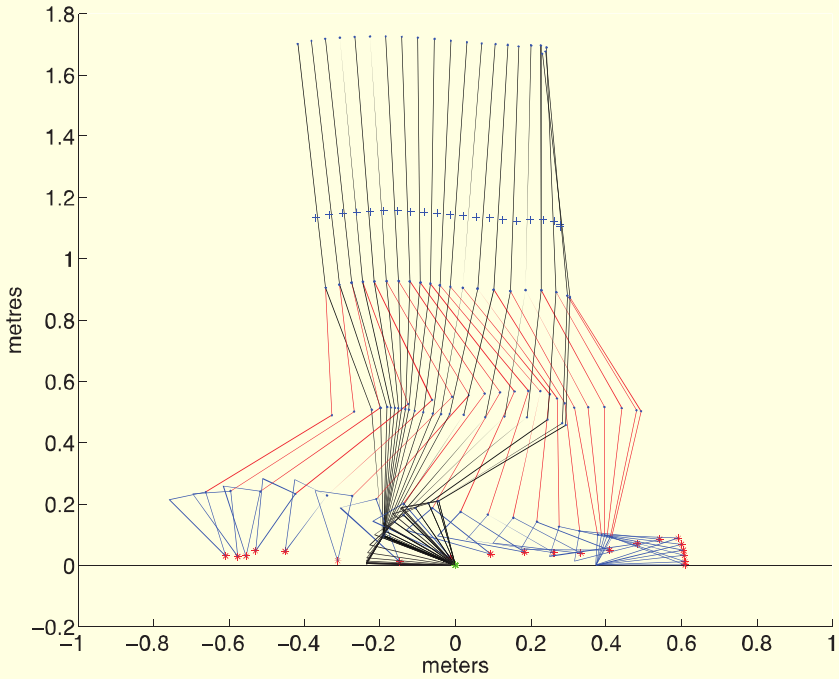



Figure 4.19: *One step walk motion - Experiment 3*

of each control were observed to be smaller in Experiment 3, while in the double support phase, as time interval was larger, knot intervals of each control were then observed to be bigger. This was to accommodate the shorter time span in single support phase and longer time span in double support phase.

DISTRIBUTION A. Approved for public release: distribution unlimited.

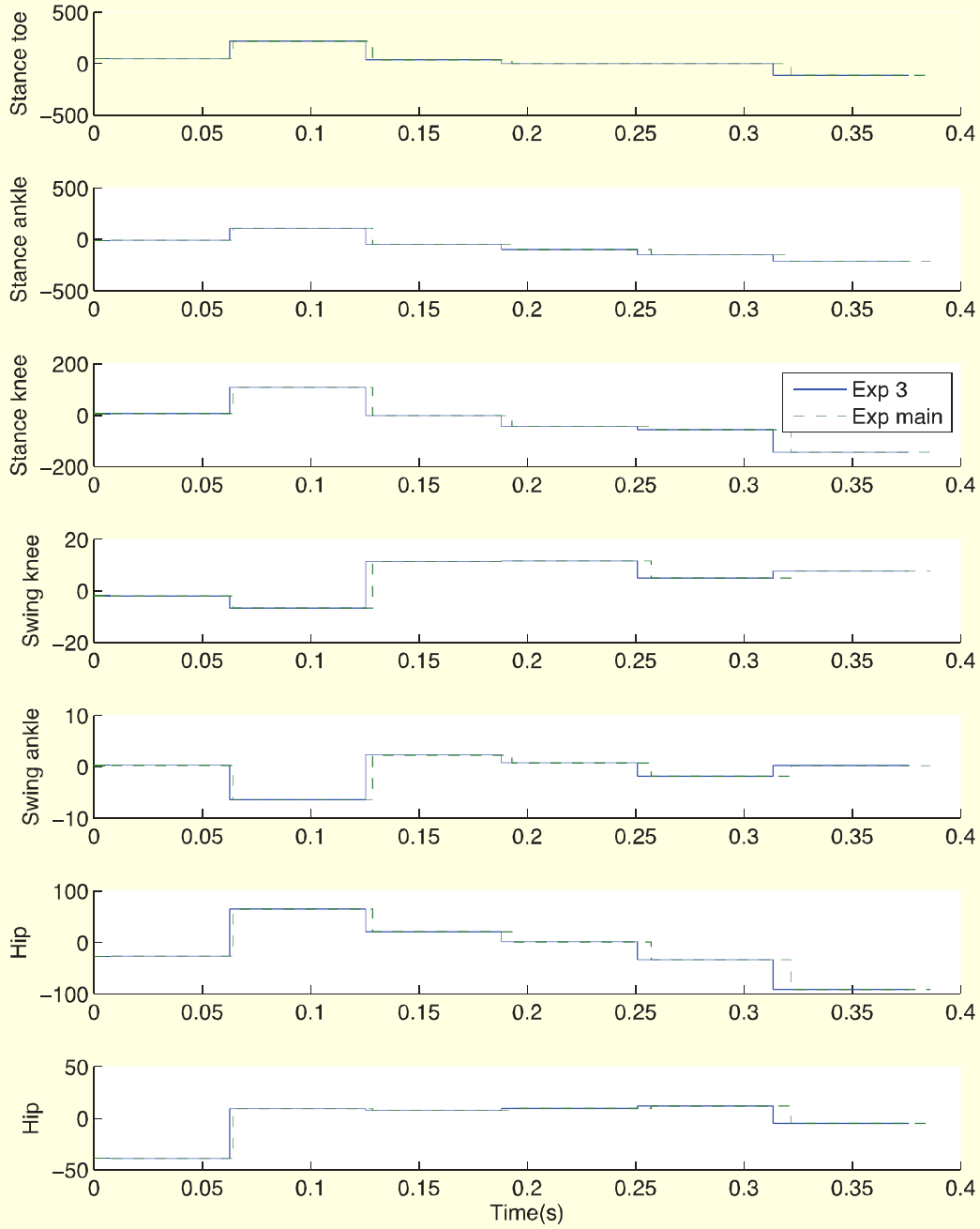



Figure 4.20: *Exp 3 versus Exp Main - optimized joint torque trajectories during single support phase*

DISTRIBUTION A. Approved for public release: distribution unlimited.

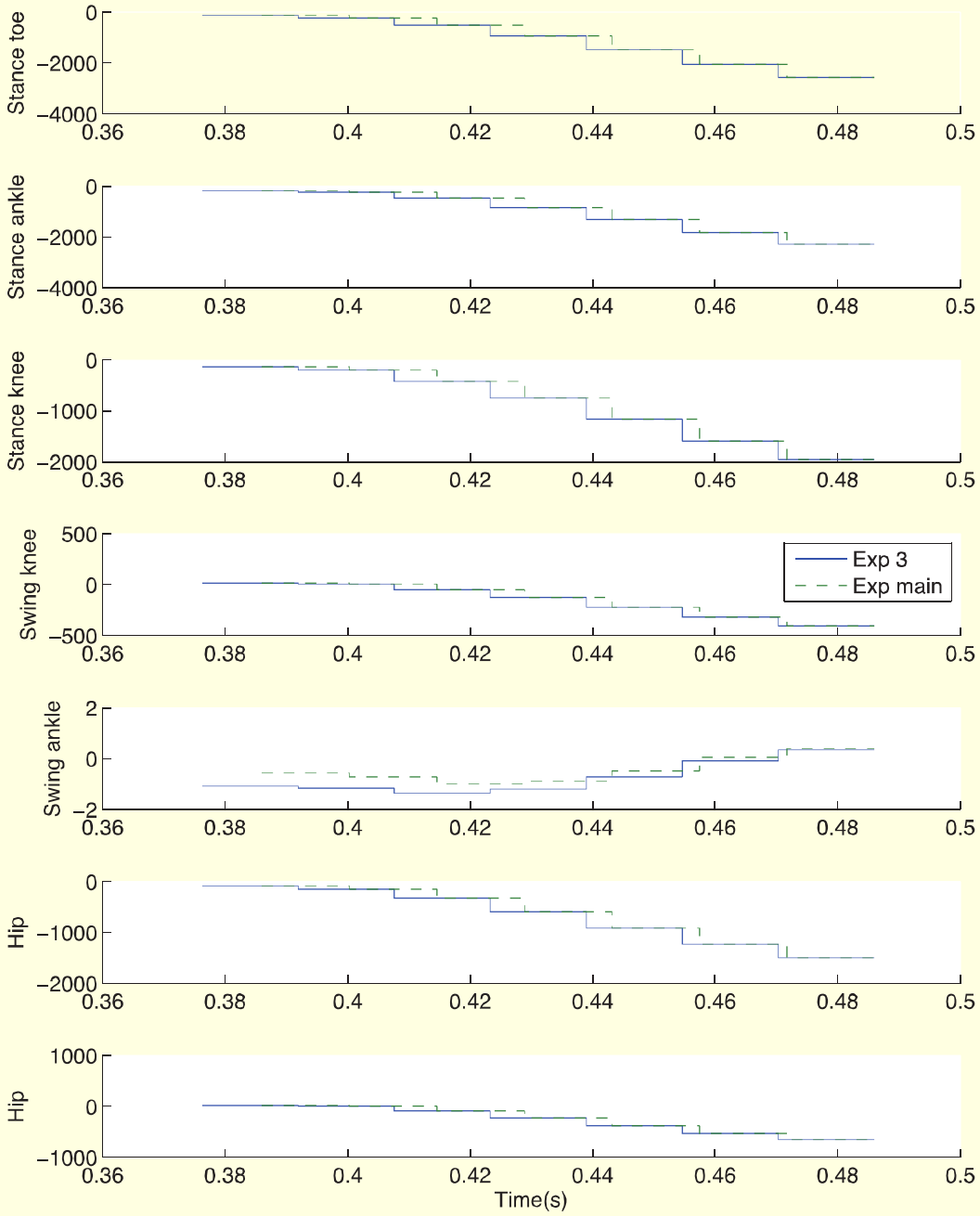



Figure 4.21: *Exp 3 versus Exp Main - optimized joint torque trajectories during double support phase*

DISTRIBUTION A. Approved for public release: distribution unlimited.

Even forces in Experiment 3, similar to joint torques, were observed to have shifted according to the changes of time interval yet maintaining a identical pattern as Experiment Main. Despite having same motion pattern, it was noticed that external vertical forces on the stance toe were smaller in Experiment 3, but greater on the swing ankle at joint 6, when comparing with Experiment Main (Figure 4.22). With larger time interval in double support phase, more time was allowed for the weight of the body to be distributed from stance foot to swing foot, as swing foot is on the ground for a longer duration. More force is acting on the swing foot as it remains longer on the ground, while stance foot prepares to lift-off.

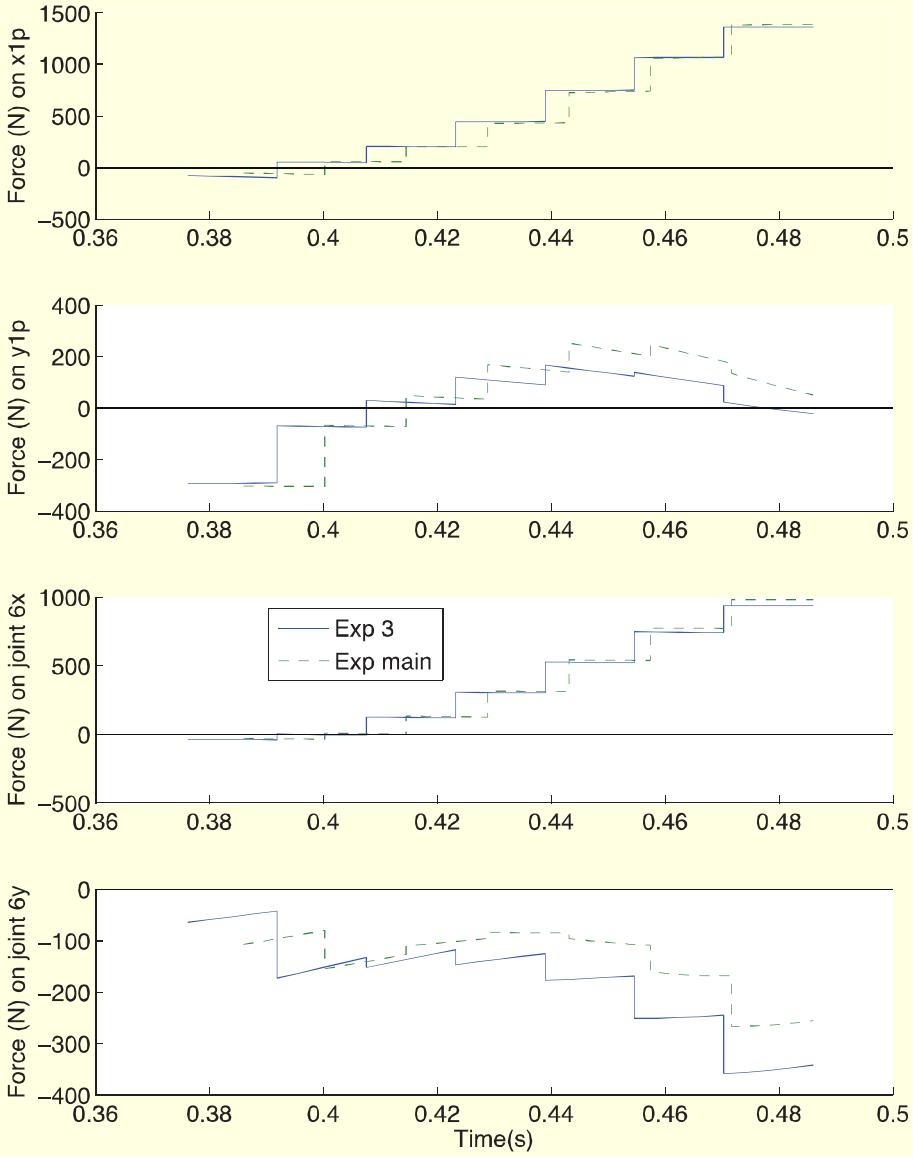



Figure 4.22: *Exp 3 versus Exp Main - External forces on joint 1 (x1p, y1p) and joint 6 (6x,6y) during double support phase*

DISTRIBUTION A. Approved for public release: distribution unlimited.

Table 4.4 provides a comparison of the optimized system parameters of Experiment 3 and Main. As the objective was to minimise step length, the lower boundary of $z_{15}$ $(= 0.61m)$ was reached in Experiment 3 to satisfy the objective. Initial angular displacements were optimized to satisfy both objective and constraints of the present experiment, and hence are different from previous models.

| Parameters | Experiment 3 | Experiment Main |
|---|---|---|
| $z_1$ | 2.7055 | 2.7055 |
| $z_2$ | 1.6370 | 1.6269 |
| $z_3$ | 1.8695 | 1.8656 |
| $z_4$ | 4.7526 | 4.5093 |
| $z_5$ | 3.7850 | 3.8011 |
| $z_6$ | 4.9598 | 4.7978 |
| $z_7$ | 1.6630 | 1.6411 |
| $z_8$ | -1.3694 | -1.2979 |
| $z_9$ | -2.3578 | -2.4389 |
| $z_{10}$ | -1.1421 | -1.3727 |
| $z_{11}$ | 1.3291 | 1.6809 |
| $z_{12}$ | 0.9549 | 1.1253 |
| $z_{13}$ | -4.1197 | -4.0952 |
| $z_{14}$ | 0.2092 | 0.3605 |
| $z_{15}$ | 0.61 | 0.7351 |

Table 4.4: *System parameters*

Figure 4.23 depicts the first and last segment of the walk cycle in Experiment 3. A periodic motion could not really be noticed in the present experiment although the only periodic constraint considered was on the trunk. A further look at angular periodicity of each segment was presented in Figure 4.24 and Table 4.5. No improvements in periodicity could be seen between Experiment 3 and Main, but Experiment 1 certainly faired better in periodicity since periodicity constraints were considered.

| Periodicity | Experiment 3 | Experiment Main | Experiment 1 |
|---|---|---|---|
| $z_6 - \theta_1(t_f) - \pi$ | 0.0383 | -0.2513 | 0.0418 |
| $z_5 - \theta_2(t_f) - \pi$ | 0.0093 | -0.0043 | -0.0533 |
| $z_4 - \theta_3(t_f) - \pi$ | 0.0676 | 0.0048 | 0.0256 |
| $z_3 - \theta_4(t_f) + \pi$ | -0.1712 | -0.2388 | -0.1196 |
| $z_2 - \theta_5(t_f) + \pi$ | 0.2519 | 0.0920 | 0.0914 |
| $z_1 - \theta_6(t_f) + \pi$ | 0.0007 | 0.0017 | 0.0369 |
| $z_7 - \theta_7(t_f)$ | 0.0004 | 0.0006 | -0.0859 |

Table 4.5: *Angular periodicity differences*

DISTRIBUTION A. Approved for public release: distribution unlimited.

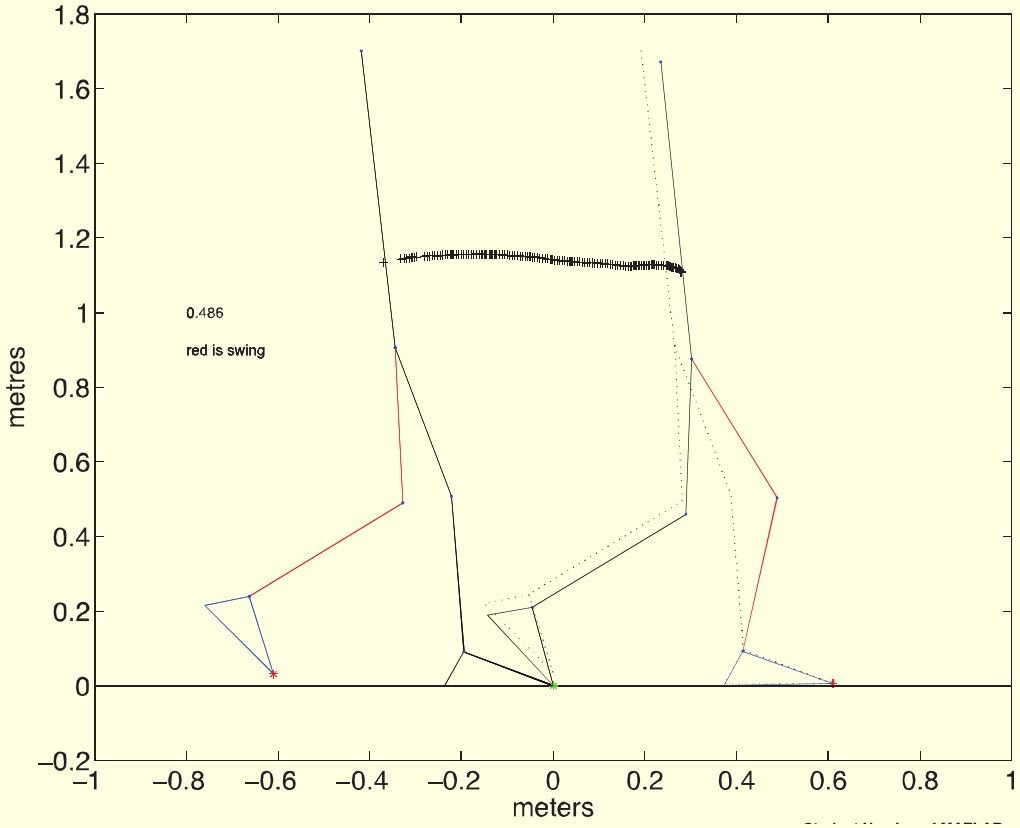



Figure 4.23: *Start and end of walk cycle with center of mass position - Experiment 3*

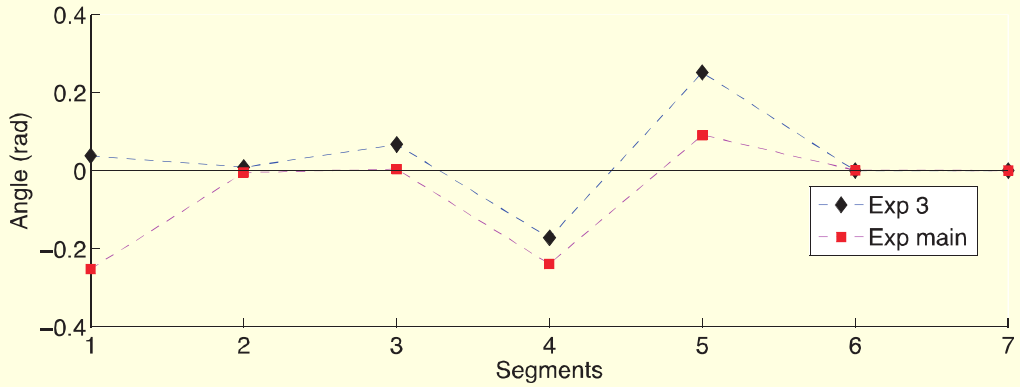



Figure 4.24: *Exp 3 versus Exp Main - Angular periodicity differences*

Results of Experiment 3 suggested that at a slower walk speed, duration of single support phase was shortened and double support phase was lengthened. Forces and torques behaved identically to Experiment Main but the knot intervals were shifted according to the changes in time interval of each phase. Unlike Experiment 2, periodic motion was no different to Experiment Main, which suggested periodicity could only be achieve at higher walking speed

DISTRIBUTION A. Approved for public release: distribution unlimited.

or when periodicity constraints were considered.

## 5 Concluding Remarks

In this paper, we have extended the mathematical method proposed in [18] to analyse human walking behaviour with the addition of periodicity constraints and at various speeds. The mathematical method proposed is able to simulate normal walking motion of a full walk cycle at different speeds, with improvement in periodicity seen with either the addition of periodicity constraints or at a faster walking speed. Periodic motion simulation was improved by our method with the addition of periodicity constraints as identified in Experiment 1 or at faster walking speed as in Experiment 2. Experiments 2 and 3 also show that our method allows velocity adjustment in modelling walk motions at different speeds. A main advantage of the mathematical method developed is its ability to model a walk motion in different, more realistic scenarios as a single process, instead of as multiple processes.

Though the method proposed here can model human periodical walking accurately and effectively, the optimal solutions are usually sensitive to external disturbances as the model proposed by us is based on an open-loop optimal control approach. To model more realistic human walking behaviours, techniques for constructing optimal feedback or robust controllers such as those in [6, 7, 17, 1, 2, 12] have to be used to yield optimal solutions that are stable and robust in the presence of internal and external disturbances. We will discuss this in future papers.

## References

[1] H. Alwardi, S. Wang, L.S. Jennings and S. Richardson, An adaptive least-squares collocation radial basis function method for the HJB equation, *Journal of Global Optimization* 52 (2012) 305–322.

[2] H. Alwardi, S. Wang and L.S. Jennings, An adaptive domain decomposition method for the Hamilton-Jacobi-Bellman equation, *Journal of Global Optimization* 56 (2013) 1361–1373.

[3] A. Brooke, D. Kendrick and A. Meeraus, *GAMS: A User's Guide, Release 2.25*, The Scientific Press, San Francisco, 1992.

[4] A. Cappozzo, T. Leo and A. Pedotti, A general computing method for the analysis of human locomotion, *Journal of Biomechanics* 8 (1975) 307–320.

[5] D. Garg, M. Patterson, W.W. Hager, A.V. Rao, D.A. Benson and G.T. Huntington, A unified framework for the numerical solution of optimal control problems using pseudospectral methods, *Automatica* 46 (2010) 184–1851.

[6] C.-S. Huang, S. Wang, and K.L. Teo. On application of an alternating direction method to Hamilton-Jacobin-Bellman equations, *J. Comp. Appl. Math.* 166 ( 2004) 153–166.

[7] C.-S. Huang, S. Wang, S.C. Chen and Z.C. Li, A radial basis collocation method for Hamilton-Jacobi-Bellman equations, *Automatica* 42 (2006) 2201–2207.

[8] L.S. Jennings, M.E. Fisher, K.L. Teo and C.J. Goh, *MISER3 optimal control software (version 3): Theory and User Manual*, Centre of Applied Dynamics and Optimization, The University of Western Australia, 2000.

**DISTRIBUTION A. Approved for public release: distribution unlimited.**

[9] M.T.H. Koh, *Optimal Performance of the Yurchenko Layout Vault*, Ph.D thesis, University of Western Australia, 2001.

[10] M.T.H. Koh and L.S. Jennings, Dynamic optimization: A solution to the inverse dynamics problem of biomechanics using MISER3, *Dynamics of Continuous, Discrete and Impulsive Systems - Series B - Applications & Algorithms* 9 ( 2002) 369–386.

[11] M.T.H. Koh and L.S. Jennings, Dynamic optimization: Inverse analysis for the Yurchenko layout vault in women's artistic gymnastics, *Journal of Biomechanics* 36 (2003) 1177–1183.

[12] M.S. Mahmoud and O. Al-Buraiki, Robust control of autonomous bicycle kinematics, *Numerical Algebra, Control and Optimization* 4 (2014)181–191.

[13] J.J. Moré and S.J. Wright, *Optimization Software Guide*, SIAM, Philadelphia, 1994.

[14] S. Onyshko and D.A. Winter, A mathematical model for the dynamics of human locomotion, *Journal of Biomechanics* 13 (1980) 361–368.

[15] M.G. Pandy, Computer modeling and simulation of human movement, *Annual Review of Biomedical Engineering* 3 (2001) 245–273.

[16] L. Ren, R.K. Jones and D. Howard, Predictive modelling of human walking over a complete gait cycle, *Journal of Biomechanics* 40 ( 2007) 1567–1574.

[17] S. Richardson, S. Wang and L.S. Jennings, A multivariate adaptive regression B-spline algorithm (BMARS) for solving a class of nonlinear optimal feedback control problems, *Automatica* 44 (2008) 1149–1155.

[18] M. Tan, L.S. Jennings and S. Wang, Analysing human walking using dynamic optimisation, in *Optimization Methods, Theory and Applications*, H. Xu, S. Wang, S.-Y. Wu (eds), Springer-Verlag, Berlin-Heidelberg, 2015, pp.1–34.

[19] K.L. Teo and L.S. Jennings, Non-linear optimal control problems with continuous state inequality constraints, *Journal of Optimization Theory and Applications* 63 (1989) 1–21.

[20] K.L. Teo, C.J. Goh and K.H. Wong, *A Unified Computational Approach to Optimal Control Problems*, Longman Scientific and Technical, Essex, England, 1991.

[21] K.L. Teo, L.S. Jennings, H.W.J. Lee and V. Rehbock, The control parametrization enhancing transform for constrained optimal control problems, *Journal Australian Mathematical Society Series B* 40 (1999) 314–335.

[22] Y. Xiang, J.S. Arora and K. Abdel-Malek, Physics-based modeling and simulation of human walking: a review of optimization-based and other approaches, *Structural and Multidisciplnary Optimization* 42 (2010) 1–23.

---

*Manuscript received 4 March 2014*
*revised 23 July 2014*
*accepted for publication 24 September 2015*

**DISTRIBUTION A. Approved for public release: distribution unlimited.**

MEIYI TAN
School of Mathematics & Statistics
The University of Western Australia
35 Stirling Highway, Crawley, WA 6009, Australia
E-mail address: joeytmy@gmail.com

LESLIE S. JENNINGS
School of Mathematics & Statistics
The University of Western Australia
35 Stirling Highway, Crawley, WA 6009, Australia
E-mail address: les.s.jennings@gmail.com

SONG WANG
Department of Mathematics & Statistics, Curtin University
GPO Box U1987, Perth, WA 6845, Australia
E-mail address: Song.Wang@curtin.edu.au

DISTRIBUTION A. Approved for public release: distribution unlimited.